

Shutao Li  
Chenglin Liu  
Yaonan Wang (Eds.)

Communications in Computer and Information Science 484

# Pattern Recognition

6th Chinese Conference, CCPR 2014  
Changsha, China, November 17–19, 2014  
Proceedings, Part II

Part 2

# Communications in Computer and Information Science

## 484

### Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),  
Rio de Janeiro, Brazil*

Phoebe Chen

*La Trobe University, Melbourne, Australia*

Alfredo Cuzzocrea

*ICAR-CNR and University of Calabria, Italy*

Xiaoyong Du

*Renmin University of China, Beijing, China*

Joaquim Filipe

*Polytechnic Institute of Setúbal, Portugal*

Orhun Kara

*TÜBİTAK BİLGE and Middle East Technical University, Turkey*

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation  
of the Russian Academy of Sciences, Russia*

Krishna M. Sivalingam

*Indian Institute of Technology Madras, India*

Dominik Ślęzak

*University of Warsaw and Infobright, Poland*

Takashi Washio

*Osaka University, Japan*

Xiaokang Yang

*Shanghai Jiao Tong University, China*

Shutao Li Chenglin Liu Yaonan Wang (Eds.)

# Pattern Recognition

6th Chinese Conference, CCPR 2014  
Changsha, China, November 17-19, 2014  
Proceedings, Part II



Springer

## Volume Editors

Shutao Li

Hunan University, Changsha, China

E-mail: shutao\_li@hnu.edu.cn

Chenglin Liu

Chinese Academy of Sciences, Beijing, China

E-mail: liucl@nlpr.ia.ac.cn

Yaonan Wang

Hunan University, Changsha, China

E-mail: yaonan@hnu.edu.cn

ISSN 1865-0929

ISBN 978-3-662-45642-2

DOI 10.1007/978-3-662-45643-9

Springer Heidelberg New York Dordrecht London

e-ISSN 1865-0937

e-ISBN 978-3-662-45643-9

Library of Congress Control Number: 2014954668

© Springer-Verlag Berlin Heidelberg 2014

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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Message from the Chairs

Welcome to the proceedings of the 2014 Chinese Conference on Pattern Recognition (CCPR 2012) held in Changsha. CCPR 2014 was the sixth in the series, following CCPR 2007 (Beijing), CCPR 2008 (Beijing), CCPR 2009 (Nanjing), CCPR 2010 (Chongqing), and CCPR 2012 (Beijing). From 2012, CCPR is held every other year, to be alternated with the Asian Conference on Pattern Recognition (ACPR), which started in 2011.

In recent years, pattern recognition has increasingly become an enabling technology for many important and often mission-critical applications such as intelligent machines, data mining, business intelligence, Internet content search, public security monitoring, etc. The emergence of big data has triggered enormous demands on pattern recognition technology. The aim of the CCPR conference is to provide a forum for scientific exchange and presentation for pattern recognition researchers in China. CCPR started in 2007 but it can be dated back to the former editions in 1980s. In China, pattern recognition research commenced in the beginning of the 1970s. The China Association for Automation organized seven editions of the National Conference on Pattern Recognition and Machine Intelligence in the 1980s.

As in the previous CCPR events, CCPR 2014 received regular submissions and it invited internationally renowned researchers to give keynote speeches. The proceedings of CCPR 2014 are published by Springer in the *Communications in Computer and Information Science* (CCIS) series. We received 216 full submissions. Each submission was reviewed by three reviewers selected from the Program Committee and other qualified researchers. Based on the reviewers' reports, 112 papers were accepted for presentation at the conference, covering diverse fields, with 14 in pattern recognition fundamentals, 14 in feature extraction and classification, 21 in computer vision, 20 in image processing and analysis, seven in video processing and analysis, ten in biometric and action recognition, seven in biomedical image analysis, nine in document and speech analysis, and ten in pattern recognition applications.

We are grateful to the keynote speakers, Prof. Bo Zhang of Tsinghua University, Prof. Jón Atli Benediktsson of University of Iceland, Prof. Chris H.Q. Ding of the University of Texas at Arlington, and Prof. Zhi-Hua Zhou of Nanjing University. Thanks are due to the authors of all submitted papers, the Program Committee members and the reviewers, and the staff of the Organizing

Committee. Without their contributions, this conference would not have been a success. We are also grateful to Springer for publishing the proceedings, and especially to Ms. Celine (Lanlan) Chang of Springer Asia for her efforts in co-ordinating the publication.

September 2014

Yaonan Wang  
Cheng-Lin Liu  
Shutao Li

## Organization

## Sponsors

National Laboratory of Pattern Recognition  
Hunan University

## Co-sponsors

Technical Committee of Pattern Recognition of Chinese Association  
for Artificial Intelligence  
Hunan Association of Automation

## Steering Committee Chair

Tieniu Tan Institute of Automation of CAS

## **Steering Committee Members**

## General Chair

Yaonan Wang Hunan University

## Program Chairs

## Organizing Chairs

Changyan Xiao Hunan University  
Xiaogang Zhang Hunan University  
Guocai Liu Hunan University  
Hongshan Yu Hunan University

## VIII Organization

### Publication Chair

Min Liu Hunan University

### Publicity Chair

Xiaoyan Liu Hunan University

### Organizing Committee

Zhigang Ling	Hunan University
Qiaokang Liang	Hunan University
Huali Li	Hunan University
Zhenjun Zhang	Hunan University
Xiaoqing Lu	Hunan University
Leyuan Fang	Hunan University
Li Zhou	Hunan University
Haiyan Zhang	Hunan University

### Technical Program Committee

Xiang Bai	Huazhong University of Science and Technology
Hong Chang	Institute of Computing Technology of CAS
Shengyong Chen	Zhejiang University of Technology
Songcan Chen	Nanjing University of Aeronautics and Astronautics
Hong Cheng	University of Electronic Science and Technology of China
Jian Cheng	Institute of Automation of CAS
Dao-Qing Dai	Sun Yat-Sen University
Junyu Dong	Ocean University of China
Leyuan Fang	Hunan University
Jianjiang Feng	Tsinghua University
Jufu Feng	Peking University
Xinbo Gao	Xidian University
Xin Geng	Southeast University
Xiaofei He	Zhejiang University
Zhaoshui He	Zhejiang University
Baogang Hu	Institute of Automation of CAS
Kaizhu Huang	Xi'an Jiaotong-Liverpool University
Yunde Jia	Beijing Institute of Technology
Lianwen Jin	South China University of Technology
Xiaoyuan Jing	Wuhan University
Jun Li	Sun Yat-Sen University
Shutao Li	Hunan University

Wu-Jun Li	Nanjing University
Xuelong Li	Xi'an Institute of Optics and Precision Mechanics of CAS
Cheng-Lin Liu	Institute of Automation of CAS
Guocai Liu	Hunan University
Min Liu	Hunan University
Qingshan Liu	Nanjing University Information Science and Technology
Wenju Liu	Institute of Automation of CAS
Yuehu Liu	Xi'an Jiaotong University
Huchuan Lu	Dalian University of Technology
Jiwen Lu	Advanced Digital Sciences Center, Singapore
Yue Lu	East China Normal University
Bin Luo	Anhui University
Zhenjiang Miao	Beijing Jiaotong University
Yanwei Pang	Tianjin University
Yu Qiao	Shenzhen Institute of Advanced Technology of CAS
He Ran	Institute of Automation of CAS
Nong Sang	Huazhong University of Science and Technology
Huanfeng Shen	Wuhan University
Jun Sun	Fujitsu R&D Center Co., LTD
Xiaoyang Tan	Nanjing University of Aeronautics and Astronautics
Jinhui Tang	Nanjing University of Science and Technology
Dacheng Tao	Hong Kong Polytechnic University
Wenbing Tao	Huazhong University of Science and Technology
Zengfu Wang	University of Science and Technology of China
Hanzi Wang	University of Adelaide
Liang Wang	Institute of Automation of CAS
Liwei Wang	Peking University
Shengjin Wang	Tsinghua University
Yaonan Wang	Hunan University
Yunhong Wang	Peking University
Yihong Wu	Institute of Automation of CAS
Shiming Xiang	Institute of Automation of CAS
Changyan Xiao	Hunan University
Xin Xu	National University of Defense Technology
Pingkun Yan	Philips Research North America
Jian Yang	Nanjing University of Science and Technology
Jian Yu	Beijing Jiaotong University
Yuan Yuan	Xi'an Institute of Optics and Precision Mechanics of CAS

Hongbin Zha	Peking University
Changshui Zhang	Tsinghua University
Daoqiang Zhang	Nanjing University of Aeronautics and Astronautics
Hongbin Zhang	Beijing University of Technology
Min-Ling Zhang	Southeast University
Jun Zhao	Institute of Automation of CAS
Wenming Zheng	Southeast University
Ping Zhong	National University of Defense Technology
Jie Zhou	Tsinghua University
Zhi-Hua Zhou	Nanjing University
Wangmeng Zuo	Harbin Institute of Technology

## Additional Reviewers

Bo Bai	Zhong Ji	Wen Lu
Jiale Cao	Sen Jia	Xiaoqiang Lu
Li Chen	Bo Jiang	Javier López-Fandiño
Mingming Chen	Wei Jiang	Longlong Ma
Pan Chen	Xiaoheng Jiang	Gabriel Martín Herández
Yuxi Chen	Ye Jiang	Liangrui Peng
Zhihui Chen	Mahdi Khodadadzadeh	Antonio Plaza
Zhenchao Cui	Taotao Lai	Javier Plaza
Qing Da	Minxian Li	Lishan Qiao
Qun Dai	Peiqiang Li	Dongwei Ren
Zhijun Dai	Qiming Li	Alim Samat
Cheng Deng	Ying Li	Xiangbo Shu
Hong Deng	Zechao Li	Benqin Song
Xiaoming Deng	Zhaoxin Li	Fengyi Song
Yongsheng Dong	Shan Liang	Yan Song
Fuqing Duan	He Lin	Dengdi Sun
Huixian Duan	Zhigang Ling	Chunna Tian
Bin Fan	Fan Liu	Nicole Tsu
Wei Fan	Lei Liu	Minghua Wan
Jianwu Fang	Li Liu	Bin Wang
Yachuang Feng	Qi Liu	Da-Han Wang
Youji Feng	Qingjie Liu	Faqiang Wang
Lianru Gao	Wei Liu	Jian Wang
Shuhang Gu	Wei Liu	Jun Wang
Lihua Guo	Yi Liu	Liang Wang
Fei He	Yu Liu	Limin Wang
Bo Hu	Luo Longrun	Liuan Wang
Sheng-Jun Huang	Guifu Lu	Peng Wang
Weilin Huang	Shujing Lu	Song Wang

Taiqing Wang	Yan Yan	Xu-Yao Zhang
Xg Wang	Aiping Yang	Yanming Zhang
Xiumei Wang	Wankou Yang	Hao Zheng
Ying Wang	Wei Yang	Kai Zheng
Zhe Wang	Xubing Yang	Liang Zheng
Ziteng Wang	Zhanlei Yang	Xiangtao Zheng
Chunpeng Wu	Cong Yao	Guoqiang Zhong
Shufu Xie	Fei Yin	Xiang-Dong Zhou
Yujie Xiong	Xiaofang Yuan	Ying Zhou
Guili Xu	Jianlong Zhang	Yu Zhou
Liang Xu	Jiaqi Zhang	Ximei Zhu
Miao Xu	Jing Zhang	Zhuotun Zhu
Xiang Xu	Mingming Zhang	Lina Zhuang
Wei Xue	Shaoquan Zhang	Pengcheng Zou
Zhaohui Xue	Xin Zhang	
Songbai Yan	Xingyi Zhang	

## Table of Contents – Part II

### Section IV: Image Processing and Analysis

Transferring Segmentation from Image to Image via Contextual Sparse Representation .....	1
<i>Shuangshuang Li, Yonghao He, Shiming Xiang, Lingfeng Wang, and Chunhong Pan</i>	
Fast Augmented Lagrangian Method for Image Smoothing with Hyper-Laplacian Gradient Prior .....	12
<i>Li Chen, Hongzhi Zhang, Dongwei Ren, David Zhang, and Wangmeng Zuo</i>	
Study on Distribution Coefficient in Regulation Services with Energy Storage System .....	22
<i>Shaojie Tan, Xinran Li, Ming Wang, Yawei Huang, Tingting Xu, and Xingtong Cheng</i>	
All-Focused Light Field Image Rendering .....	32
<i>Rumin Zhang, Yu Ruan, Dijun Liu, and Zhang Youguang</i>	
Hyperspectral Image Unmixing Based on Sparse and Minimum Volume Constrained Nonnegative Matrix Factorization .....	44
<i>Denggang Li, Shutao Li, and Huali Li</i>	
An Adaptive Harris Corner Detection Algorithm for Image Mosaic .....	53
<i>Haixia Pan, Yanxiang Zhang, Chunlong Li, and Huafeng Wang</i>	
A Study of Ancient Ceramics Verification Based on Vision Methods .....	63
<i>Yunqi Tang, Jianwei Ding, and Wei Guo</i>	
A Two-Stage Blind Image Color Correction Using Color Cast Estimation .....	72
<i>Dawei Zhu, Li Chen, Jing Tian, and Xiaotong Huang</i>	
Encoding Optimization Using Nearest Neighbor Descriptor .....	81
<i>Muhammad Rauf, Yongzhen Huang, and Liang Wang</i>	
Multi-modal Image Fusion with KNN Matting .....	89
<i>Xia Zhang, Hui Lin, Xudong Kang, and Shutao Li</i>	
A Two-Step Adaptive Descreening Method for Scanned Halftone Image .....	97
<i>Fei Chen, Shutao Li, Le Xu, Bin Sun, and Jun Sun</i>	

Compressive Sensing Multi-focus Image Fusion . . . . .	107
<i>Fang Cheng, Bin Yang, and Zhiwei Huang</i>	
Pan-Sharpening Based on Improvement of Panchromatic Image to Minimize Spectral Distortion . . . . .	117
<i>Akbi Abdelkrim, Zhaoxiang Zhang, and Qingjie Liu</i>	
Combining SIFT and Individual Entropy Correlation Coefficient for Image Registration . . . . .	128
<i>Gan Liu, Shengyong Chen, Xiaolong Zhou, Xiaoyan Wang, Qiu Guan, and Hui Yu</i>	
Spectral Fidelity Analysis of Compressed Sensing Reconstruction Hyperspectral Remote Sensing Image Based on Wavelet Transformation . . . . .	138
<i>Yi Ma, Jie Zhang, and Ni An</i>	
A Fast Algorithm for Image Defogging . . . . .	149
<i>Xiaoyan He, Jianxu Mao, Zewen Liu, Jiujiang Zhou, and Yajing Hua</i>	
A New Image Structural Similarity Metric Based on K-L Transform . . . . .	159
<i>Cheng Jiang, Fen Xiao, and Xiaobo He</i>	
A New Restoration Algorithm for Single Image Defogging . . . . .	169
<i>Fan Guo, Hui Peng, and Jin Tang</i>	
An Improved Laparoscopic Image Registration Algorithm Based on Sift . . . . .	179
<i>Jiujiang Zhou, Jianxu Mao, and Xiaoyan He</i>	
Application of Image Processing Techniques in Infrared Detection of Faulty Insulators . . . . .	189
<i>Yefan Wu, Jiangang Yao, Tangbing Li, Peng Fu, Wei Liao, and Mi Zhang</i>	

## Section V: Video Processing and Analysis

Finding the Accurate Natural Contour of Non-rigid Objects in Video . . . . .	199
<i>Gaoxuan Ying, Sheng Liu, and Yiting Jin</i>	
An Improved Multipitch Tracking Algorithm with Empirical Mode Decomposition . . . . .	209
<i>Wei Jiang, Wenju Liu, Yingwei Tan, and Shan Liang</i>	
Robust Appearance Learning for Object Tracking in Challenging Scenes . . . . .	218
<i>Jianwei Ding, Yunqi Tang, Huawei Tian, and Yongzhen Huang</i>	

Vehicle Recognition for Surveillance Video Using Sparse Coding .....	228
<i>Shirong Zeng, Xin Niu, and Yong Dou</i>	

Video Smoke Detection Based on the Optical Properties .....	235
<i>Yingjing Wu and Ying Hu</i>	

Discovery of the Topical Object in Commercial Video: A Sparse Coding Method .....	245
<i>Yunhui Liu, Huaping Liu, and Fuchun Sun</i>	

Study the Moving Objects Extraction and Tracking Used the Moving Blobs Method in Fisheye Image .....	255
<i>Jianhui Wu, Guoyun Zhang, Shuai Yuan, Longyuan Guo, and Mengxia Tan</i>	

## **Section VI: Biometric and Action Recognition**

A Non-negative Low Rank and Sparse Model for Action Recognition ...	266
<i>Biyun Sheng, Wankou Yang, Baochang Zhang, and Changyin Sun</i>	

Extreme Learning Machine Based Hand Posture Recognition in Color-Depth Image .....	276
<i>Zhen Zhou, Shutao Li, and Bin Sun</i>	

Real-Time Human Detection Based on Optimized Integrated Channel Features .....	286
<i>Jifeng Shen, Xin Zuo, Wankou Yang, and Guohai Liu</i>	

Facial Feature Extraction Based on Robust PCA and Histogram .....	296
<i>Xiao Luan and Weisheng Li</i>	

Multimodal Finger Feature Fusion and Recognition Based on Delaunay Triangular Granulation .....	303
<i>Jinjin Peng, Yanan Li, Ruimei Li, Guimin Jia, and Jinfeng Yang</i>	

Robust Face Recognition via Facial Disguise Learning .....	311
<i>Meng Yang and Linlin Shen</i>	

A Static Hand Gesture Recognition Algorithm Based on Krawtchouk Moments .....	321
<i>Shuping Liu, Yu Liu, Jun Yu, and Zengfu Wang</i>	

Face Recognition in the Wild by Mining Frequent Feature Itemset .....	331
<i>Yuzhuo Wang, Hong Cheng, Yali Zheng, and Lu Yang</i>	

Single-Sample Face Recognition via Fusion Variant Dictionary .....	341
<i>Ying Tai, Jian Yang, Jianjun Qian, and Yu Chen</i>	

Supervised Kernel Construction for Unsupervised PCA on Face Recognition .....	351
<i>Yang Zhao, Wen-Sheng Chen, Binbin Pan, and Bo Chen</i>	

## Section VII: Biomedical Image Analysis

Medical Image Clustering Based on Improved Particle Swarm Optimization and Expectation Maximization Algorithm.....	360
<i>Zheng Tang, Yuqing Song, and Zhe Liu</i>	
Medical Image Fusion by Combining Nonsubsampled Contourlet Transform and Sparse Representation.....	372
<i>Yu Liu, Shuping Liu, and Zengfu Wang</i>	
Automated Segmentation and Tracking of SAM Cells .....	382
<i>Min Liu and Peng Xiang</i>	
Automatic Estimation of Muscle Thickness in Ultrasound Images Based on Revoting Hough Transform (RVHT) .....	392
<i>Jianhao Tan, Xiaolong Li, Wentao Zhang, Yaoqin Xie, and Yongjin Zhou</i>	

Influence of Scan Duration on the Reliability of Resting-State fMRI Regional Homogeneity .....	402
<i>Xiaotang Li, Jiansong Zhou, and Xiaoyan Liu</i>	

A Global Eigenvalue-Driven Balanced Deconvolution Approach for Network Direct-Coupling Analysis.....	409
<i>Haiping Sun and Hongbin Shen</i>	

Sequence-Based Prediction of Protein-Protein Binding Residues in Alpha-Helical Membrane Proteins .....	419
<i>Feng Xiao and Hongbin Shen</i>	

## Section VIII: Document and Speech Analysis

Robust Voice Activity Detection Using the Combination of Short-Term and Long-Term Spectral Patterns .....	428
<i>Yingwei Tan and Wenju Liu</i>	
Speech Emotion Recognition Based on Coiflet Wavelet Packet Cepstral Coefficients .....	436
<i>Yongming Huang, Ao Wu, Guobao Zhang, and Yue Li</i>	
Text Detection in Natural Scene Images Leveraging Context Information .....	444
<i>Runmin Wang, Nong Sang, Changxin Gao, Xiaoqin Kuang, and Jun Xiang</i>	

Adaptive Local Receptive Field Convolutional Neural Networks for Handwritten Chinese Character Recognition .....	455
<i>Li Chen, Chunpeng Wu, Wei Fan, Jun Sun, and Naoi Satoshi</i>	
Character Segmentation for Classical Mongolian Words in Historical Documents .....	464
<i>Xiangdong Su, Guanglai Gao, Weihua Wang, Feilong Bao, and Hongxi Wei</i>	
MCDF Based On-Line Handwritten Character Recognition for Total Uyghur Character Forms .....	474
<i>Askar Hamdulla, Wujiahemaiti Simayi, Mayire Ibrayim, and Dilmurat Tursun</i>	
Natural Scene Text Image Compression Using JPEG2000 ROI Coding .....	481
<i>Yuanping Zhu and Li Song</i>	
Off-Line Uyghur Handwritten Signature Verification Based on Combined Features .....	491
<i>Kurban Ubul, Tuergen Yibulayin, and Alimjan Aysa</i>	
Off-Line Signature Verification Based on Local Structural Pattern Distribution Features .....	499
<i>Wen Jing, MoHan Chen, and JiaXin Ren</i>	
<b>Section IX: Pattern Recognition Applications</b>	
Coordination of Electric Vehicles Charging to Maximize Economic Benefits .....	508
<i>Yongwang Zhang, Haoming Yu, Chun Huang, Wei Zhao, and Min Luo</i>	
Traffic Sign Recognition Using Perturbation Method .....	518
<i>Linlin Huang and Fei Yin</i>	
A Novel Two-Stage Multi-objective Ant Colony Optimization Approach for Epistasis Learning .....	528
<i>Pengjie Jing and Hongbin Shen</i>	
Hydraulic Excavators Recognition Based on Inverse "V" Feature of Mechanical Arm .....	536
<i>Wenming Yang, Dedi Li, Daren Sun, and Qingmin Liao</i>	
Real-Time Traffic Sign Detection via Color Probability Model and Integral Channel Features .....	545
<i>Yi Yang and Fuchao Wu</i>	

## XVIII Table of Contents – Part II

Study of Charging Station Short-Term Load Forecast Based on Wavelet Neural Networks for Electric Buses . . . . .	555
<i>Lei Zhang, Chun Huang, and Haoming Yu</i>	
The Layout Optimization of Charging Stations for Electric Vehicles Based on the Chaos Particle Swarm Algorithm . . . . .	565
<i>Zhenghui Zhang, Qingxiu Huang, Chun Huang, Xiuguang Yuan, and Dewei Zhang</i>	
An Improved Feature Weighted Fuzzy Clustering Algorithm with Its Application in Short-Term Prediction of Wind Power . . . . .	575
<i>Xinkun Wang, Diansheng Luo, and Hongying He</i>	
Charging Load Forecasting for Electric Vehicles Based on Fuzzy Inference . . . . .	585
<i>Jingwei Yang, Diansheng Luo, Shuang Yang, and Shiyu Hu</i>	
Security Event Classification Method for Fiber-optic Perimeter Security System Based on Optimized Incremental Support Vector Machine . . . . .	595
<i>Lu Liu, Wei Sun, Yan Zhou, Yuan Li, Jun Zheng, and Botao Ren</i>	
<b>Author Index . . . . .</b>	<b>605</b>

# Table of Contents – Part I

## Section I: Fundamentals of Pattern Recognition

A Nonlinear Classifier Based on Factorization Machines Model . . . . .	1
<i>Xiaolong Liu, Yanming Zhang, and Chenglin Liu</i>	
Training Deep Belief Network with Sparse Hidden Units . . . . .	11
<i>Zhen Hu, Wenzheng Hu, and Changshui Zhang</i>	
The Research of Matching Area Selection Criterion for Gravity Gradient Aided Navigation . . . . .	21
<i>Kaihan Li, Ling Xiong, Long Cheng, and Jie Ma</i>	
A Manifold Learning Fusion Algorithm Based on Distance and Angle Preservation . . . . .	31
<i>Yanchun Gu, Defeng Zhang, Zhengming Ma, and Guo Niu</i>	
Application of Modified Teaching-Learning Algorithm in Coordination Optimization of TCSC and SVC . . . . .	44
<i>Liwu Xiao, Qianlong Zhu, Canbing Li, Yijia Cao, Yi Tan, and Lijuan Li</i>	
Multi-task Sparse Gaussian Processes with Improved Multi-task Sparsity Regularization . . . . .	54
<i>Jiang Zhu and Shiliang Sun</i>	
Short-Term Load Forecasting of LSSVM Based on Improved PSO Algorithm . . . . .	63
<i>Qianhui Gong, Wenjun Lu, Wenlong Gong, and Xueting Wang</i>	
Blob Detection with the Determinant of the Hessian . . . . .	72
<i>Xiaopeng Xu</i>	
A Study on Layer Connection Strategies in Stacked Convolutional Deep Belief Networks . . . . .	81
<i>Lei Guo, Shijie Li, Xin Niu, and Yong Dou</i>	
Fault Diagnosis for Distribution Networks Based on Fuzzy Information Fusion . . . . .	91
<i>Fangrong Wu, Minfang Peng, Mingjun Qi, Liang Zhu, Hua Leng, Yi Su, Qiang Zhong, and Hu Tan</i>	
Kernel-Distance Target Alignment . . . . .	101
<i>Peiyan Wang and Cai Dongfeng</i>	

A LLE-Based HMM Applied to the Prediction of Kiln Coal Feeding Trend .....	111
<i>Yunlong Liu and Zhang Xiaogang</i>	

Improved Margin Sampling for Active Learning .....	120
<i>Jin Zhou and Shiliang Sun</i>	

Research on the Ant Colony Optimization Fuzzy Neural Network Control Algorithm for ABS .....	130
<i>Changping Wang and Ling Wang</i>	

## **Section II: Feature Extraction and Classification**

Schatten p-Norm Based Matrix Regression Model for Image Classification .....	140
<i>Lei Luo, Jian Yang, Jinhui Chen, and Yicheng Gao</i>	

Hyperspectral Image Classification by Exploiting the Spectral-Spatial Correlations in the Sparse Coefficients .....	151
<i>Dan Liu, Shutao Li, and Leyuan Fang</i>	

Spectral-Spatial Hyperspectral Image Classification Using Superpixel and Extreme Learning Machines .....	159
<i>Wuhui Duan, Shutao Li, and Leyuan Fang</i>	

Visual Tracking with Weighted Online Feature Selection .....	168
<i>Yu Tang, Zhigang Ling, Jiancheng Li, and Lu Bai</i>	

A System of Image Aesthetic Classification and Evaluation Using Cloud Computing .....	183
<i>Weining Wang, Jiancong Liu, Weijian Zhao, and Jiachang Li</i>	

Image Feature Extraction via Graph Embedding Regularized Projective Non-negative Matrix Factorization .....	196
<i>Haishun Du, Qingpu Hu, Xudong Zhang, and Yandong Hou</i>	

Sparse Manifold Preserving for Hyperspectral Image Classification .....	210
<i>Hong Huang, Fulin Luo, Jiamin Liu, and Zehzhong Ma</i>	

Hyperspectral Image Classification Using Local Collaborative Representation .....	219
<i>Yishu Peng, Yunhui Yan, Wenjie Zhu, and Jiuliang Zhao</i>	

Simplified Constraints Rank-SVM for Multi-label Classification .....	229
<i>Jiarong Wang, Jun Feng, Xia Sun, Su-Shing Chen, and Bo Chen</i>	

Semi-supervised Image Classification Learning Based on Random Feature Subspace .....	237
<i>Liu Li, Zhang Huaxiang, Hu Xiaojun, and Sun Feifei</i>	

An Improved Multi-label Classification Ensemble Learning Algorithm.....	243
<i>Zhongliang Fu, Lili Wang, and Danpu Zhang</i>	

An Improved Sparse Representation De-noising for Keeping Structural Features .....	253
<i>Zhi Cui</i>	

A SVM Method Trained by Improved Particle Swarm Optimization for Image Classification .....	263
<i>Qifeng Qian, Hao Gao, and Baoyun Wang</i>	

Sparsity Based Feature Extraction for Kernel Minimum Squared Error .....	273
<i>Jiang Jiang, Xi Chen, Haitao Gan, and Nong Sang</i>	

### **Section III: Computer Vision**

Saliency Detection Based on Spread Pattern and Manifold Ranking .....	283
<i>Yan Huang, Keren Fu, Lixiu Yao, Qiang Wu, and Jie Yang</i>	

A Structural Constraint Based Dual Camera Model .....	293
<i>Xinzhou Li, Yuehu Liu, Shaozhuo Zhai, and Zhichao Cui</i>	

Hough Voting with Distinctive Mid-Level Parts for Object Detection ...	305
<i>Xiaoqin Kuang, Nong Sang, Feifei Chen, Runmin Wang, and Changxin Gao</i>	

A Segmentation Based Change Detection Method for High Resolution Remote Sensing Image .....	314
<i>Lin Wu, Zhaoxiang Zhang, Yunhong Wang, and Qingjie Liu</i>	

Eye Localization Based on Multi-Channel Correlation Filter Bank.....	325
<i>Rui Yang, Shiming Ge, Kaixuan Xie, and Shuixian Chen</i>	

Person Re-identification by Cascade-Iterative Ranking .....	335
<i>Xiangyu Wang, Feng Chen, and Yu Liu</i>	

Stereo Camera Based Real-Time Local Path-Planning for Mobile Robots .....	345
<i>Huanqing Yang, Jianhua Zhang, and Shenyong Chen</i>	

A Tracking Method with Structural Local Mean and Local Standard Deviation Appearance Model .....	355
<i>Dawei Yang, Yang Cong, Yandong Tang, and Yulian Li</i>	

Hough-RANSAC: A Fast and Robust Method for Rejecting Mismatches .....	363
<i>Hongxia Gao, Jianhe Xie, Yueming Hu, and Ze Yang</i>	

Partial Static Objects Based Scan Registration on the Campus . . . . .	371
<i>Chongyang Wei, Shuangyin Shang, Tao Wu, and Hao Fu</i>	
Quasi-Orthorectified Panorama Generation Based on Affine Model from Terrain UAV Images . . . . .	381
<i>Yuchong Li</i>	
Shape Recognition by Combining Contour and Skeleton into a Mid-Level Representation . . . . .	391
<i>Wei Shen, Xinggang Wang, Cong Yao, and Xiang Bai</i>	
Visual Texture Perception with Feature Learning Models and Deep Architectures . . . . .	401
<i>Yuchen Zheng, Guoqiang Zhong, Jun Liu, Xiaoxu Cai, and Junyu Dong</i>	
Objects Detection Method by Learning Lifted Wavelet Filters . . . . .	411
<i>Aireti Abulikemu, Aliya Yushan, Turghunjan Abdurakim Turki, and Abdurusul Osman</i>	
A Fast Straight-Line Growing Algorithm for Sheet-Counting with Stacked-Paper Images . . . . .	418
<i>ZhenXiao Gang, Shuo Yang, and Changyan Xiao</i>	
Automatic Labanotation Generation Based on Human Motion Capture Data . . . . .	426
<i>Hao Guo, Zhenjiang Miao, Feiyue Zhu, Gang Zhang, and Song Li</i>	
Self-organizing Map-Based Object Tracking with Saliency Map and K-Means Segmentation . . . . .	436
<i>Yuanding Zhang, Yuanyan Tang, Bin Fang, Zhaowei Shang, and C.Y. Suen</i>	
Superpixel-Based Global Optimization Method for Stereo Disparity Estimation . . . . .	445
<i>Haiqiang Jin, Sheng Liu, Shaobo Zhang, and Gaoxuan Ying</i>	
Two-Stage Saliency Detection Based on Continuous CRF and Sparse Coding . . . . .	455
<i>Qiyang Zhao, Weibo Li, Fan Wang, and Baolin Yin</i>	
Continuous Energy Minimization Based Multi-target Tracking . . . . .	464
<i>Zhe Shi, Songhao Zhu, Wei Sun, and Baoyun Wang</i>	
<b>Author Index . . . . .</b>	<b>475</b>

# Transferring Segmentation from Image to Image via Contextual Sparse Representation

Shuangshuang Li, Yonghao He, Shiming Xiang, Lingfeng Wang, and Chunhong Pan

NLPR, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China

**Abstract.** It is still a fundamental task to segment objects out from diverse background. To tackle this task, we propose a transferring segmentation framework, which aims to automatically segment new images when a single segmented example is given. Our segmentation approach is developed under the observation that some regions of foreground and background are often very similar but rarely share similar contextual information. To this end, we propose to construct a contextual dictionary by incorporating neighboring information as context. The segmentation task is finally accomplished in way of supervised classification via sparse representation with the constructed contextual dictionary. Experimental results on diverse natural images demonstrate that the proposed method achieves favorable results in both visual quality and accuracy.

**Keywords:** Image Segmentation, Contextual Dictionary, Sparse Representation, Superpixel.

## 1 Introduction

Image segmentation is one of the most fundamental tasks in image processing. The task of image segmentation is to partition an image into regions with homogenous visual appearance. Although there are many thoughtful attempts for decades, it is still very difficult to develop a general approach that can be applied to diverse natural images.

In the literature, most previous segmentation methods are developed based on information related to edges, shapes, colors and textures. Later, graph based approaches are considered to model and represent natural images, among which the most typical approach is the normalized cuts [1] based on spectral graph theory. These approaches attempt to generate segmentations with regions of pixels sharing the same visual characteristics. However, they fail to yield satisfactory results which can describe the semantic concepts of images. The degradation of these automatic segmentation approaches is due to two difficulties [2]. On the low level, it is difficult to describe the visual elements including colors, textures, and shapes. On the high level, no generic rules can be used to group the visual patterns into meaningful regions.

Technically, it is necessary to introduce supervised information to generate better segmentations. Now different forms have been designed to utilize the supervised information well, typically including strongly supervised segmentation [3] and interactive segmentation [2,4,5,6]. Most existing strongly supervised segmentation algorithms require a large number of training images as well as the corresponding ground truths to train segmentation models. However, this requirement is difficult to meet for real-world



**Fig. 1.** Description of transferring segmentation. (a) The example image and its ground truth. (b) The image to be segmented.

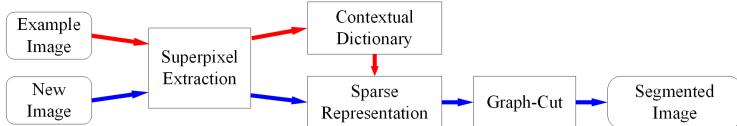
applications, as manually annotating the ground truths is costly. Interactive segmentation methods require the user to label the foreground and background. While in most cases, these algorithms only deal with a single image at a time and manual interaction is required for each image. When there are a large number of images to be segmented, the process of interaction is very time-consuming and labor-intensive.

Recently, cosegmentation has been considered into image segmentation [7,8]. The aim of cosegmentation is to simultaneously segment multiple interrelated images into foreground and background. The hypothesis of cosegmentation lies in that images should contain similar foreground. With this hypothesis, all the regions with similar visual appearance will be segmented as foreground. Thus, the task is developed without supplying the ground truths of the training images. However, when images also contain similar background, these methods may fail to recognize the segments which belong to the foreground.

In this paper, we explore another possibility to segment an image automatically in terms of transferring segmentation, which refers to transferring the segmentation information from an example image to new images. Essentially, our task is addressed into a supervised segmentation framework, where the example image is regarded as a training one.

Fig. 1 illustrates the segmentation setting and our task. Suppose we are given an image and its segmented result, which are taken as a pair of examples (see Fig .1a). Our goal is to automatically segment new similar images (see Fig. 1b) according to the given example. Thus, it is different from the aforementioned strongly supervised segmentation or interactive segmentation, since it does not require user interaction or a large training set. Additionally, it also differs from the cosegmentation approaches in which they do not need segmented images for training.

Our technical outline is illustrated in Fig. 2 . Images are first over-segmented into superpixels that are treated as training samples. Specifically, the foreground and background contextual dictionaries are constructed via the contextual information of the training samples (see subsection 3.1). The contextual information is explored and exploited from the superpixels of the training image. For a new image, the dictionaries will be employed to obtain the contextual sparse representation (see subsection 3.2). Then the reconstruction errors on the foreground and background dictionaries will be calculated respectively, and these errors are then used to classify the superpixels as foreground or background. Finally, graph-cut approach is used to improve the segmentation accuracy (see subsection 3.3).



**Fig. 2.** Flowchart of our algorithm. Red lines stand for the training process, while blue lines denote the segmentation procedure.

## 2 Brief Review of Sparse Representation

Sparse representation has been widely used in computer vision and pattern recognition, such as image restoration, denoising, classification [9,10] and recognition [11,12]. Let  $\|\cdot\|_p$  be the  $\ell^p$ -norm for  $p=0,1,2$ . Given an over-complete dictionary  $\mathbf{D} \in \mathbb{R}^{n \times K}$  which contains  $K$  atoms (codewords), the sparse representation of a signal  $\mathbf{y} \in \mathbb{R}^n$  can be formulated as follows:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0, \quad \text{subject to } \mathbf{y} = \mathbf{D}\boldsymbol{\alpha}, \quad (1)$$

where  $\boldsymbol{\alpha} \in \mathbb{R}^K$  contains the representation coefficients.

However, (1) is an NP-hard problem. Recent research often relaxes it as the following  $\ell^1$ -minimization problem:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \mu\|\boldsymbol{\alpha}\|_1. \quad (2)$$

In (2), the first term is the reconstruction error, and the second term restricts the sparsity of  $\boldsymbol{\alpha}$ . The positive parameter  $\mu$  balances the sparsity and the reconstruction error.

## 3 Contextual Sparse Representation for Transferring Segmentation

In this section, we will present our transferring segmentation approach. As mentioned previously, our task will be addressed under the framework of sparse representation. To this end, the dictionaries will be constructed by utilizing the contextual information that is explored from superpixels of the training image. Then, the image to be segmented will be represented sparsely with the constructed dictionaries from which the foreground will be inferred. Finally, the graph-cut (GC) approach is used to refine the results.

### 3.1 The Construction of Contextual Dictionaries

In sparse representation, a key issue is to construct the dictionary. According to the proposal for face recognition given by Wright et al. [11], one way is to select patches from the training images and take each patch as an atom of the dictionary. Another way is to learn a dictionary [13] from the sampled patches. In practice, we notice some patches of foreground and background have similar visual appearance. If we directly

construct or learn a dictionary from them, some atoms of the foreground dictionary and the background dictionary could be nearly identical to each other. This will decrease the discriminative power of the dictionary, and thus may cause unsatisfactory segmentation. To remedy this drawback, we consider to combine the contextual information into the dictionary construction. Here superpixels [14,15] are employed to extract the contextual information. In our work, Simple Linear Iterative Clustering (SLIC) [14,16] is used to extract superpixels. SLIC runs fast and can preserve most of the boundaries of the input images. Images are segmented into about 800-1200 superpixels, since too coarse segmentations may fail to preserve the image boundaries and too fine segmentations will lead to inadequate contextual information as well as high complexity.

Given the training image, the SLIC method is employed to segment it into superpixels. Suppose we obtain  $N$  superpixels. For clarity, we denote the  $i$ -th superpixel by  $\mathbf{x}_i$  ( $i = 1, 2, \dots, N$ ), which is described by using the mean RGB color and the normalized RGB histogram of the pixels in this superpixel. Since the size of the superpixels is small (about 150-200 pixels), we use eight bins to describe each bit plane. Thus, there are totally 512 ( $8 \times 8 \times 8$ ) bins in the histogram. As a result, each  $\mathbf{x}_i$  is a vector in  $\mathbb{R}^{512}$ . Consequently, we take each superpixel as a sample, and collect them in  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ .

Then, we construct two dictionaries respectively from the foreground and background regions of the training image, and denote them by  $\mathbf{D} = [\mathbf{D}^f, \mathbf{D}^b]$ , where  $\mathbf{D}^f$  is the foreground dictionary and  $\mathbf{D}^b$  refers to the background dictionary. Specifically, for each  $\mathbf{x}_i$ , its neighboring samples including itself are first picked out, which are collected in neighborhood  $\mathcal{N}_i$ , namely,  $\mathcal{N}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{n_i}$ . Here  $n_i$  is the number of the samples in  $\mathcal{N}_i$ ,  $\mathbf{x}_{i,j}$  is a sample in  $\mathcal{X}$ , and the subscript  $i,j$  denotes its index in  $\{1, 2, \dots, N\}$ .

Note that each  $\mathcal{N}_i$  may contain different number of neighboring samples, which have different sizes and region shapes of pixels. Let  $a_{i,j}$  be the area of the region covered by the sample  $\mathbf{x}_{i,j}$ . Thus, we further employ the area of each sample as weights, and calculate the weighted average  $\mathbf{x}_i^c$  as follows:

$$\mathbf{x}_i^c = \frac{\sum_{j=1}^{n_i} a_{i,j} \mathbf{x}_{i,j}}{\sum_{j=1}^{n_i} a_{i,j}}. \quad (3)$$

When calculating  $\mathbf{x}_i^c$  in (3), for each  $\mathbf{x}_{i,j}$ , we only consider the features corresponding to the histogram. Now, each atom in the contextual dictionary is constructed by concatenating  $\mathbf{x}_i$  and  $\mathbf{x}_i^c$  together:

$$\mathbf{d}_i = [\mathbf{x}_i^T, \beta(\mathbf{x}_i^c)^T]^T, \quad (4)$$

where  $\beta$  is a positive parameter, which controls the importance of the contextual information in the dictionary.

Finally, the foreground and background contextual dictionaries are constructed by grouping the atoms as follows:

$$\mathbf{D}^f = \{\mathbf{d}_i | \mathbf{x}_i \text{ belongs to the foreground}\}, \quad (5)$$

$$\mathbf{D}^b = \{\mathbf{d}_j | \mathbf{x}_j \text{ belongs to the background}\}. \quad (6)$$

The atoms are constructed in this way because it is less possible for foreground samples and background samples to share similar neighboring information. This will help enhance the discriminative power of the dictionaries  $\mathbf{D}^f$  and  $\mathbf{D}^b$ .

### 3.2 Contextual Sparse Representation for the New Image

Given a new image  $\mathcal{I}$  to be segmented, we employ the SLIC method to partition it into superpixels. For each superpixel  $\mathbf{y}$  in  $\mathcal{I}$ , all of its neighboring superpixels are first identified and the averaged histogram is then calculated like the way described in (3). Denoting its averaged histogram by  $\mathbf{y}^c$ , now our task is to sparsely represent  $[\mathbf{y}^T, \beta(\mathbf{y}^c)^T]^T$  by using the concatenation of the two dictionaries  $\mathbf{D}^f$  and  $\mathbf{D}^b$ . This task can be described as an optimization problem as follows:

$$\min_{\alpha^f, \alpha^b} \left\| \begin{bmatrix} \mathbf{y} \\ \beta \mathbf{y}^c \end{bmatrix} - [\mathbf{D}^f, \mathbf{D}^b] \begin{bmatrix} \alpha^f \\ \alpha^b \end{bmatrix} \right\|_2^2 + \gamma (\|\alpha^f\|_1 + \|\alpha^b\|_1), \quad (7)$$

where  $\alpha^f$  and  $\alpha^b$  are the sparse coefficients with respect to  $\mathbf{D}^f$  and  $\mathbf{D}^b$ , and  $\gamma$  is a positive number, which controls the sparsity of the coefficients. We use homotopy [17,18] to obtain the contextual sparse representation (CSR) in (7).

Based on  $\alpha^f$  and  $\alpha^b$ , the reconstruction errors can be calculated through:

$$e^k = \left\| \begin{bmatrix} \mathbf{y} \\ \beta \mathbf{y}^c \end{bmatrix} - \mathbf{D}^k \alpha^k \right\|_2^2, \quad \text{where } k \in \{f, b\}. \quad (8)$$

Then, the corresponding label  $l$  of  $\mathbf{y}$  is determined by:

$$l = \arg \min_k \{e^k\}, \quad \text{where } k \in \{f, b\}. \quad (9)$$

In (9),  $l = f$  stands for  $\mathbf{y}$  belonging to the foreground, while  $l = b$  for the background.

Finally, we explain why the concatenation of  $\mathbf{D}^f$  and  $\mathbf{D}^b$  is used to represent the vector  $[\mathbf{y}^T, \beta(\mathbf{y}^c)^T]^T$ . Actually, this vector can be linearly reconstructed with  $\mathbf{D}^f$  and  $\mathbf{D}^b$  respectively, from each of which one can obtain a reconstruction error. However, this vector may be both well reconstructed with these dictionaries, thus the discriminative power of the errors will be largely reduced. In contrast, when we put the atoms of  $\mathbf{D}^f$  and  $\mathbf{D}^b$  together like (7), the most similar atoms to this vector will be selected with the sparse representation. This will cause a large gap between  $e^f$  and  $e^b$ , and finally facilitate the classification.

### 3.3 Refinement by Graph-Cut Framework

Note that the performance of our proposed contextual sparse representation (CSR) in subsection 3.2 will be used to classify each of the superpixels one-by-one. This treatment does not consider the similarity between superpixels. In addition, some superpixels may be misclassified and cause noisy segmentation. Considering the spatial continuity in natural images, we introduce smoothness constraints to reduce the misclassification rate. Here the popular graph-cut (GC) approach provides a natural way for incorporating such constraints [4]. After refining the results of CSR through GC, the proposed framework is called CSRGC.

Let  $\mathcal{L} = \{l_i\}_{i=1}^M$  denote the labels of the testing samples (superpixels) of the new image  $\mathcal{I}$ , where  $M$  is the number of samples. Then we can construct a graph by taking

each sample as a node. According to the graph-cut approach, the segmentation can be achieved by minimizing the following Gibbs energy  $E(\mathcal{L})$  [4]:

$$E(\mathcal{L}) = \sum_{i=1}^M \Psi(l_i) + \lambda \sum_{\langle i,j \rangle} \Phi(l_i, l_j), \quad (10)$$

where  $\langle i, j \rangle$  stands for an edge connecting adjacent samples. Parameter  $\lambda$  is a smoothness regularization coefficient, which is allowed to take values in the range  $[0.001, 0.1]$ .  $\Psi(l_i)$  is the likelihood energy indicating the cost of labeling the  $i$ -th sample with  $l_i$ , which is defined as:

$$\begin{aligned} \Psi(l_i = f) &= -\log \frac{e_i^b}{e_i^f + e_i^b}, \\ \Psi(l_i = b) &= -\log \frac{e_i^f}{e_i^f + e_i^b}, \end{aligned} \quad (11)$$

where  $e_i^f$  and  $e_i^b$  are the reconstruction errors calculated by (8). Let  $\mathbf{c}_i$  be the mean RGB color of the  $i$ -th sample. In (10),  $\Phi(l_i, l_j)$  is the smoothness term defined as:

$$\Phi(l_i, l_j) = \frac{1}{\|\mathbf{c}_i - \mathbf{c}_j\|_2 + \delta} [l_i \neq l_j], \quad (12)$$

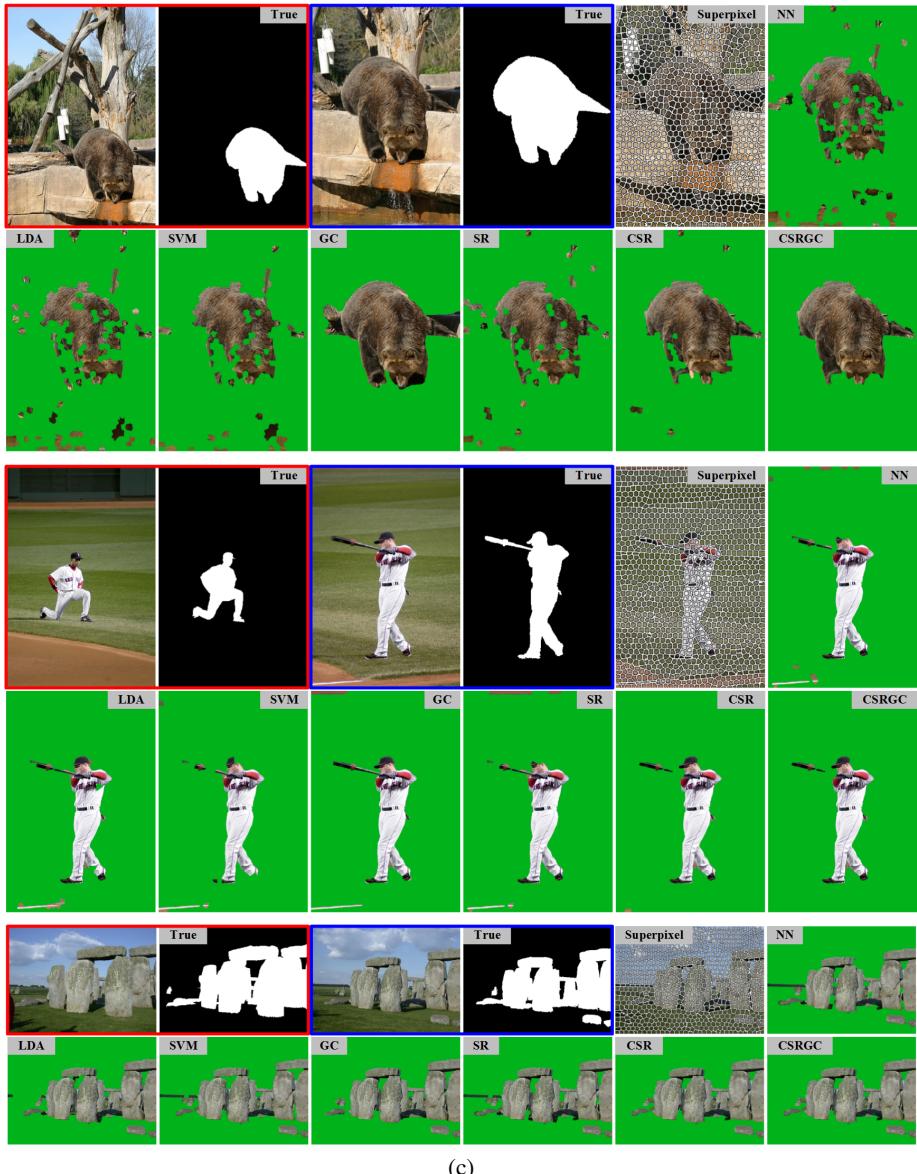
where  $\delta = 0.0001$  is a small positive constant, and  $[\cdot]$  is the zero-one indicator function.

## 4 Experiments

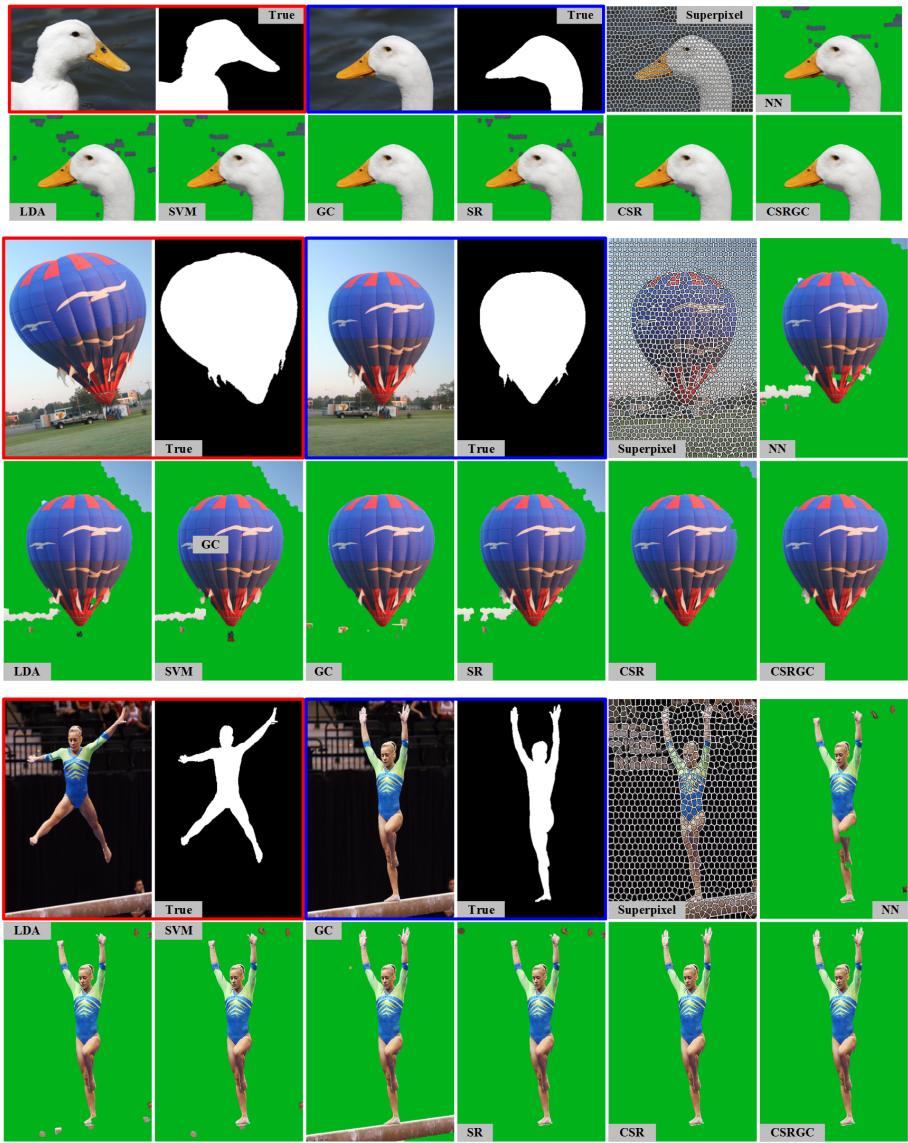
The images in our experiments are downloaded from CMU-Cornell iCoseg dataset [19,20] and Flower Datasets [21]. Since image segmentation is essentially a classification problem, the proposed method is compared with support vector machine (SVM), nearest neighbor (NN) classifier and linear discriminant analysis with NN classifier (LDA). Furthermore, we also compare the performance of CSRG with GC, which uses the example image to obtain Gaussian mixture models for the foreground and background. We use libsvm [22] to implement SVM with the Gaussian kernel function. The parameter related to Gaussian kernel is set to be 0.03, which is selected from  $[0.01, 10]$  and performs well on the test images in Fig. 3 and Fig. 4. The regularization parameter is set to be 100, since poor results may be generated if it is very small.

In our method, the parameter  $\beta$  controls the importance of contextual information used in the dictionary. The parameter  $\beta$  is experimentally set to be 2.8. It is worth noting that when  $\beta = 0$ , no contextual information is considered and the method is degraded into sparse representation (SR). To show the advantage of contextual information, we also display the results of SR. Parameter  $\gamma$  is set to be 0.01.

Fig. 3 and Fig. 4 displays the segmentation results on image pairs. When the contextual information is used (i.e. CSR), the results are better than that of SR. On the average, our method CSRG performs best. Please pay attention that some images are rather difficult for image segmentation, since some regions of foreground and background share similar colors. And our method is insensitive to rotation (e.g. Fig. 4b), scale change (e.g. Fig. 3a), view point change (e.g. Fig. 4a) and foreground/background change to some extent (e.g. Fig. 3b). To further compare these algorithms, a quantitative comparison



**Fig. 3.** Segmentation results of the images from CMU-Cornell iCoseg dataset [20]. The images are scaled for arrangement and the background is filled with green in the results. The original sizes of the test images are  $500 \times 333$ ,  $500 \times 333$  and  $375 \times 500$ , respectively.



(c)

**Fig. 4.** Segmentation results of the images from CMU-Cornell iCoseg dataset [20]. The images are scaled for arrangement and the background is filled with green in the results. The example image of (b) is rotated by 10 degrees to test the robustness of rotation. The original sizes of the test images are  $333 \times 500$ ,  $500 \times 333$  and  $500 \times 400$ , respectively.

between these algorithms is given in Tabel 1. The intersection-over-union accuracy on the pixel level is used to evaluate the results. It is calculated as  $TP/(TP+FP+FN)$ , where  $TP$  is true positive,  $FP$  is false positive and  $FN$  is false negative. Table 1 shows that our algorithm achieves the highest accuracy among all the algorithms.

**Table 1.** Accuracy of the algorithms

Images	Accuracy						
	NN	LDA	SVM	GC	SR	CSR	CSRGC
Fig. 3a	0.619	0.513	0.620	0.838	0.710	0.851	<b>0.912</b>
Fig. 3b	0.852	0.874	0.862	0.906	0.840	0.934	<b>0.940</b>
Fig. 3c	0.912	0.936	0.903	<b>0.962</b>	0.921	0.941	0.941
Fig. 4a	0.869	0.813	0.866	0.957	0.875	0.983	<b>0.990</b>
Fig. 4b	0.908	0.809	0.821	0.969	0.816	0.964	<b>0.982</b>
Fig. 4c	0.877	0.890	0.906	0.606	0.911	0.972	<b>0.972</b>
Avg.	0.840	0.806	0.830	0.873	0.846	0.941	<b>0.956</b>



**Fig. 5.** Segmenting multiple images. The images are from Flower Datasets [21]. The example image and its ground truth are in the fist column. From the second to the last columns are the results with CSRGC.

In addition, Fig. 5 shows that our algorithm CSRGC can automatically segment multiple new arriving images based on one example image.

## 5 Conclusion

In this paper, we propose a new framework of transferring the segmentation from an example image to new arriving images. With this framework, the segmentation results are obtained by supervised classification, which is implemented by the sparse representation on a contextual dictionary. We address that the contextual information is very important for reducing the ambiguity of foreground and background when they share

similar colors. Thus the contextual information is exploited to improve the representational and discriminative power of the dictionary. Experimental results on diverse natural images illustrate the effectiveness of our method, and show the robustness to changes of object scale, rotation and view point.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China under Grants 61272331, 61375024, 61331018, and 91338202.

## References

1. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE TPAMI* 22(8), 888–905 (2000)
2. Xiang, S., Pan, C., Nie, F., et al.: Interactive image segmentation with multiple linear reconstructions in Windows. *IEEE Transactions on Multimedia* 13(2), 342–352 (2011)
3. Packer, B., Gould, S., Koller, D.: A unified contour-pixel model for figure-ground segmentation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part V. LNCS*, vol. 6315, pp. 338–351. Springer, Heidelberg (2010)
4. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: *ICCV*, pp. 105–112 (2001)
5. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)* 23(3), 309–314 (2004)
6. Xiang, S., Nie, F., Zhang, C.: Semi-supervised classification via local spline regression. *IEEE TPAMI* 32(11), 2039–2053 (2010)
7. Rother, C., Minka, T., Blake, A., et al.: Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrf's. In: *CVPR*, pp. 993–1000 (2006)
8. Dai, J., Wu, Y., Zhou, J., et al.: Cosegmentation and Cosketch by Unsupervised Learning. In: *ICCV*, pp. 1305–1312 (2013)
9. Yuan, H., Lu, Y., et al.: Sparse representation using contextual information for hyperspectral image classification. In: *IEEE International Conference on Cybernetics*, pp. 138–143 (2013)
10. Fang, L., Li, S., et al.: Spectral-Spatial hyperspectral image classification via multiscale adaptive sparse representation. *IEEE Transactions on Geoscience and Remote Sensing* 52(12), 7738–7749 (2014)
11. Wright, J., Yang, A.Y., Ganesh, A., et al.: Robust face recognition via sparse representation. *IEEE TPAMI* 22(8), 210–227 (2009)
12. Zhang, Q., Li, B.: Discriminative K-SVD for dictionary learning in face recognition. In: *CVPR*, pp. 2691–2698 (2010)
13. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54(11), 4311–4322 (2006)
14. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI* 34(11), 2274–2282 (2012)
15. Xiang, S., Pan, C., Nie, F., Zhang, C.: Turbopixel segmentation using eigen-images. *TIP* 19(11), 3024–3034 (2010)
16. Image and Visual Representation Group,  
<http://ivrg.epfl.ch/research/superpixels>
17. Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* 20(3), 389–403 (2000)

18. L1 Homotopy: A MATLAB Toolbox for Homotopy Algorithms in L1 Norm Minimization Problems, <http://users.ece.gatech.edu/~sasif/homotopy/index.html>
19. Batra, D., Kowdle, A., Parikh, D., et al.: Icoseg: Interactive co-segmentation with intelligent scribble guidance. In: CVPR, pp. 3169–3176 (2010)
20. Advanced Multimedia Processing (AMP) Lab, <http://chenlab.ece.cornell.edu/projects/touch-coseg/>
21. National Center for Biotechnology Information, <http://www.robots.ox.ac.uk/~vgg/data/flowers/>
22. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM TIST 2(3), 27:1–27:27 (2011)

# Fast Augmented Lagrangian Method for Image Smoothing with Hyper-Laplacian Gradient Prior

Li Chen<sup>1</sup>, Hongzhi Zhang<sup>1</sup>, Dongwei Ren<sup>1</sup>, David Zhang<sup>1, 2</sup>,  
and Wangmeng Zuo<sup>1</sup>

<sup>1</sup> Computational Perception and Cognition Center, School of Computer Science  
and Technology, Harbin Institute of Technology, Harbin, 150001, China

{lichen.cs.hit, zhanghz0451, rendongweihit, cswmzuo}@gmail.com

<sup>2</sup> Biometrics Research Centre, Department of Computing, The Hong Kong Polytechnic  
University, Hung Hom, Kowloon, Hong Kong  
csdzhang@comp.polyu.edu.hk

**Abstract.** As a fundamental tool,  $L_0$  gradient smoothing has found a flurry of applications. Inspired by the progress of research on hyper-Laplacian prior, we propose a novel model, corresponding to  $L_p$ -norm of gradients, for image smoothing, which can better maintain the general structure, whereas diminishing insignificant texture and impulse noise-like highlights. Algorithmically, we use augmented Lagrangian method (ALM) to efficiently solve the optimization problem. Thanks to the fast convergence rate of ALM, the speed of the proposed method is much faster than the  $L_0$  gradient method. We apply the proposed method to natural image smoothing, cartoon artifacts removal, and tongue image segmentation, and the experimental results validate the performance of the proposed algorithm.

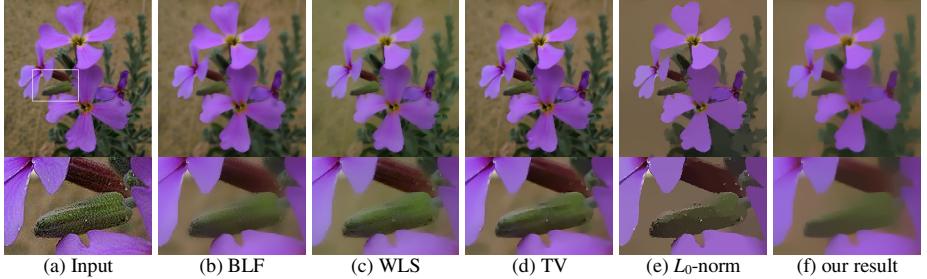
**Keywords:** Image smoothing, augmented Lagrangian method, hyper-Laplacian gradient prior.

## 1 Introduction

Noise and blur usually are inevitable in real world images. In many image processing applications, e.g., edge detection [1], object segmentation [2], etc., it is necessary to enhance the well-structured components, whereas suppressing noise and unnecessary texture. As a conventional approach for image denoising and enhancement, image smoothing is a well-studied problem with quite a number of methods in literatures, such as bilateral filtering (BLF) [3], weighted least squares (WLS) [4], total variation (TV) [5],  $L_0$ -norm smoothing [6].

Recent studies on natural image statistics have shown that heavy-tailed image gradient distribution [7] is an effective prior and can be well modeled by hyper-Laplacian with  $0 < p < 1$  [8, 9, 10], which has been applied in several applications, e.g., image deburring, leading to superior results. At the same time, many efforts have been devoted to research on the  $L_p$  optimization, e.g., iteratively reweighted least squares (IRLS) [11, 12], iteratively reweighted  $L_1$ -minimization (IRL1) [13]. Recently, Zuo *et al.* proposed a generalized iterative shrinkage algorithm (GISA) [14].

In this paper, we propose to model image prior as an  $L_p$ -norm of gradients, which can better maintain the general structure, whereas suppressing insignificant texture. Furthermore, we adopt the augmented Lagrangian method (ALM) to efficiently solve the optimization problem, resulting in an ALM-Lp algorithm. Compared with  $L_0$ -norm smoothing, ALM-Lp algorithm is more effective in removing more highlights and avoiding color distortion, leading to a better visual perception.



**Fig. 1.** The smoothing results on image *pflower*

Fig. 1 compares the image smoothing results obtained using different approaches on a natural image *pflower*. From Fig. 1(b)-(d), one can see that BLF, WLS, and TV are valuable in wiping off noises, but are limited in removing detailed textures and preserving the salient edges. And the result of  $L_0$  smoothing is much better in the above aspects, but has color distortion to some extent. Besides, it performs poor in removing impulse noise-like highlights in the close-up. Compared with the competing methods, our result can obtain better smoothing result, which is effective in preserving the salient edges and removing noises and detailed textures. Moreover, thanks to the fast convergence rate of ALM, ALM-Lp is faster than  $L_0$  smoothing.

The remainder of this paper is organized as follows. Section 2 presents a brief review of  $L_0$  smoothing and the introduction of generalized shrinkage/thresholding (GST) algorithm. In Section 3, we present the  $L_p$ -norm smoothing model and the ALM-based optimization. Section 4 provides the experiment results. Finally, we end this paper with some concluding remarks in Section 5.

## 2 Related Work and Prerequisites

In this section, we first briefly review the  $L_0$  smoothing method, and then summarize the generalized shrinkage/thresholding (GST) function used in this paper.

### 2.1 $L_0$ Smoothing Method

Xu *et al.* [6] proposed a smoothing model based  $L_0$  norm of gradients:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \mu \|\mathbf{Dx}\|_0, \quad (1)$$

where  $\mathbf{y}$  and  $\mathbf{x}$  denote the input and smoothed image, respectively,  $\mu$  is a smoothing weight,  $\|\bullet\|_0$  denotes the  $L_0$  norm that counts the number of non-zero entries, and  $\mathbf{D} = [\mathbf{D}_h, \mathbf{D}_v]$  denotes the gradient operator that includes the horizontal component  $\mathbf{D}_h$  and vertical component  $\mathbf{D}_v$ , respectively. Then, Xu *et al.* presented an alternating optimization strategy with half-quadratic splitting to tackle this problem.

The proposed ALM-Lp method is different from  $L_0$  smoothing at two aspects. First, we adopt the  $L_p$ -norm gradient prior while  $L_0$  smoothing adopts the  $L_0$ -norm gradient prior. Previous studies [7] indicated that image gradients typically follow hyper-Laplacian distribution with  $0.5 \leq p \leq 0.8$ , which makes ALM-Lp more suitable for image smoothing. Second, ALM-Lp uses the augmented Lagrangian method (ALM) to solve the optimization, and is more efficient.

## 2.2 The Generalized Shrinkage/Thresholding (GST) Function

$L_p$ -norm minimization problem is the key of many sparse coding problems. Here, we discuss the simplest  $L_p$ -minimization problem as follows,

$$\min_x \frac{1}{2} (x - y)^2 + \lambda |x|^p. \quad (2)$$

Zuo *et al.* [14] introduced the generalized shrinkage/thresholding (GST) function which is a generalization of the soft-thresholding operator. Compared with the existing solvers for  $L_p$ -norm minimization, e.g., IRLS, LUT [7], and ITM- $L_p$  [15], GST is very efficient and converges to the correct solution. The GST function is defined as,

$$T_p^{GST}(y; \lambda) = \begin{cases} 0, & \text{if } |y| \leq \tau_p^{GST}(\lambda), \\ \text{sgn}(y) S_p^{GST}(|y|; \lambda), & \text{else} \end{cases}, \quad (3)$$

where  $\tau_p^{GST}(\lambda) = (2\lambda(1-p))^{\frac{1}{2-p}} + \lambda p (2\lambda(1-p))^{\frac{p-1}{2-p}}$  which stands for the thresholding value. Generally, if  $|y| \leq \tau_p^{GST}(\lambda)$ , the generalized soft-thresholding operator uses the thresholding rule to assign  $T_p^{GST}(y; \lambda)$  to 0; otherwise, uses the shrinkage rule to assign  $T_p^{GST}(y; \lambda)$  to  $\text{sgn}(y) S_p^{GST}(|y|; \lambda)$ .  $S_p^{GST}(|y|; \lambda)$  can be obtained by iteratively performing the following operation :

$$S_p^{GST}(y; \lambda) = y - \lambda p (S_p^{GST}(y; \lambda))^{p-1}. \quad (4)$$

Then the overall algorithm was summarized as,

$$T_p^{GST}(y; \lambda) = GST(y, \lambda, p, J), \quad (5)$$

where  $J$  is the number of iterations, generally 1 or 2.

### 3 Model and Algorithm

In this section, we first present the proposed model, formulated as an  $L_p$  norm minimization problem. For the non-convexity of the  $L_p$  norm, it is not trivial to directly optimize the proposed model. Thus, by adopting the variable splitting strategy, we employ augmented Lagrangian method (ALM) to solve it efficiently.

The smoothing model with hyper-Laplacian gradient prior is formulated as,

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mu \|\mathbf{Dx}\|_p^p. \quad (6)$$

By introducing auxiliary variable  $\mathbf{d} = \mathbf{Dx}$ , problem (6) can be reformulated as,

$$\min_{\mathbf{x}, \mathbf{d}} \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mu \|\mathbf{d}\|_p^p \quad s.t. \quad \mathbf{d} = \mathbf{Dx}, \quad (7)$$

where  $\mathbf{d} = [\mathbf{d}_h^T, \mathbf{d}_v^T]^T$  with  $\mathbf{d}_h = \mathbf{D}_h \mathbf{x}, \mathbf{d}_v = \mathbf{D}_v \mathbf{x}$ . Then the augmented Lagrangian (AL) function of the problem in Eq. (7) is defined as,

$$\mathcal{L} = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mu \|\mathbf{d}\|_p^p + \boldsymbol{\lambda}^T (\mathbf{Dx} - \mathbf{d}) + \frac{\delta}{2} \|\mathbf{Dx} - \mathbf{d}\|_2^2, \quad (8)$$

where  $\boldsymbol{\lambda}$  is the Lagrangian vector, and  $\delta$  is a positive penalty parameter. With minor algebra, the AL function can be rewritten as,

$$\mathcal{L}(\mathbf{x}, \mathbf{d}, \mathbf{q}, \delta) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mu \|\mathbf{d}\|_p^p + \frac{\delta}{2} \|\mathbf{Dx} - \mathbf{d} + \mathbf{q}\|_2^2, \quad (9)$$

where  $\mathbf{q} = \boldsymbol{\lambda}/\delta$  and  $\mathbf{q} = [\mathbf{q}_h^T, \mathbf{q}_v^T]^T$ . The AL function can be optimized using the alternating direction method of multipliers, i.e., updating  $\mathbf{x}$  while  $\mathbf{d}$  is fixed, and vice-versa.

#### 3.1 The $\mathbf{x}$ -subproblem

By fixing the variable  $\mathbf{d}$ , the subproblem w.r.t.  $\mathbf{x}$  is a quadratic optimization problem and the solution to  $\mathbf{x}$  is

$$\mathbf{x} = \left( \mathbf{1} + \delta \left( \mathbf{D}_h^T \mathbf{D}_h + \mathbf{D}_v^T \mathbf{D}_v \right) \right)^{-1} \left( \mathbf{y} + \delta \left( \mathbf{D}_h^T (\mathbf{d}_h - \mathbf{q}_h) + \mathbf{D}_v^T (\mathbf{d}_v - \mathbf{q}_v) \right) \right), \quad (10)$$

where  $\mathbf{1}$  means the delta function, and the inversion operation can be efficiently computed in the Fourier domain. Assuming circular boundary conditions, we can apply 2D fast Fourier transform (FFT) which diagonalizes the derivative operators, and the close-form solution of  $\mathbf{x}$  is

$$\mathbf{x} = \mathcal{F}^{-1} \frac{\mathcal{F}(\mathbf{y}) + \delta \mathcal{F}(\mathbf{D}_h^T (\mathbf{d}_h - \mathbf{q}_h)) + \mathcal{F}(\mathbf{D}_v^T (\mathbf{d}_v - \mathbf{q}_v))}{\mathcal{F}(\mathbf{1} + \delta (\mathbf{D}_h^T \mathbf{D}_h + \mathbf{D}_v^T \mathbf{D}_v))}, \quad (11)$$

where  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  denote FFT and inverse FFT, respectively. The plus, subtract, multiplication and division are all component-wise.

### 3.2 The $\mathbf{d}$ -subproblem

The  $\mathbf{d}$  subproblem can be formulated as,

$$\hat{\mathbf{d}} = \arg \min_{\mathbf{d}} \frac{1}{2} \|\mathbf{d} - (\mathbf{D}\mathbf{x} + \mathbf{q})\|_2^2 + \mu/\delta \|\mathbf{d}\|_p^p, \quad (12)$$

which can be solved by GST [14],

$$\mathbf{d}_h = GST(\mathbf{D}_h \mathbf{x} + \mathbf{q}_h, \mu/\delta, p, J), \mathbf{d}_v = GST(\mathbf{D}_v \mathbf{x} + \mathbf{q}_v, \mu/\delta, p, J). \quad (13)$$

Once obtaining  $\mathbf{x}$  and  $\mathbf{d}$ ,  $\mathbf{q}$  can be updated as follows

$$\mathbf{q}^{(t+1)} = \mathbf{q}^{(t)} + \mathbf{D}\mathbf{x}^{(t+1)} - \mathbf{d}^{(t+1)}. \quad (14)$$

For the updating of penalty parameter  $\delta$ , we adopt the adaptive updating strategy proposed by Lin *et al.* [16] to accelerate the convergence speed,

$$\delta^{(t+1)} = \min(\delta_{\max}, \rho \delta^{(t)}), \quad (15)$$

where  $\delta_{\max}$  is the upper bound of  $\delta$ , and the values of  $\rho$  is defined as,

$$\rho = \begin{cases} \rho_0, & \text{if } \delta \|\mathbf{d}^{(t+1)} - \mathbf{d}^{(t)}\| / \|\mathbf{D}\mathbf{x}^{(t+1)}\| < \varepsilon \\ 1, & \text{otherwise} \end{cases}. \quad (16)$$

where  $\rho_0 > 1$  is a positive constant.

---

#### Algorithm 1. ALM-Lp

---

1. **Input:** image  $\mathbf{y}$ , smoothing weight  $\mu$ , parameters  $\delta_0, \delta_{\max}$
  2. **Initialize:**  $\mathbf{x} = \mathbf{y}, \mathbf{d}^{(0)}, \mathbf{q}^{(0)}, t=0$
  3. **Precompute:**  $\mathcal{F}(\mathbf{y}), \mathcal{F}(\mathbf{H}) = \mathcal{F}(\mathbf{I} + \delta(\mathbf{D}_h^T \mathbf{D}_h + \mathbf{D}_v^T \mathbf{D}_v))$
  4. **While** not converged
  5.  $\mathbf{s}_h = \mathbf{d}_h^{(t)} - \mathbf{q}_h^{(t)}, \mathbf{s}_v = \mathbf{d}_v^{(t)} - \mathbf{q}_v^{(t)}$
  6.  $\mathbf{x}^{(t+1)} = \mathcal{F}^{-1} \left( \left( \mathcal{F}(\mathbf{y}) + \delta \left( \mathcal{F}(\mathbf{D}_h^T \mathbf{s}_h + \mathbf{D}_v^T \mathbf{s}_v) \right) \right) \otimes \mathcal{F}(\mathbf{H}) \right)$
  7.  $\mathbf{d}^{(t+1)} = GST(\mathbf{D}\mathbf{x}^{(t)} + \mathbf{q}^{(t)}, \mu/\delta, p, J)$
  8.  $\mathbf{q}^{(t+1)} = \mathbf{q}^{(t)} + \mathbf{D}\mathbf{x}^{(t+1)} - \mathbf{d}^{(t+1)}$
  9. Update  $\delta^{(t+1)}$
  10.  $t=t+1$
  11. **End while**
-

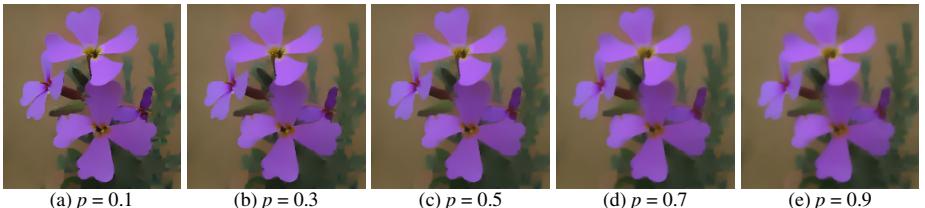
We summarize the proposed ALM-Lp algorithm in Algorithm 1. Since the Fourier transform of  $\mathcal{F}(\mathbf{y})$ ,  $\mathcal{F}(\mathbf{D}_h^T)$ ,  $\mathcal{F}(\mathbf{D}_v^T)$ , and  $\mathcal{F}(\mathbf{I} + \delta(\mathbf{D}_h^T \mathbf{D}_h + \mathbf{D}_v^T \mathbf{D}_v))$  can be pre-computed, the proposed algorithm requires 2 FFT operations per iteration. ALM-Lp involves several parameters, we empirically fix  $\delta_0 = 4\mu$ ,  $\delta_{\max} = 10^5$  and  $\epsilon = 10^{-2}$  in our experiments.

## 4 Experimental Results

In this section, we evaluate the proposed ALM-Lp method. We first evaluate the smoothing results of ALM-Lp obtained using different  $p$  values, and further compare ALM-Lp with several state-of-the-art smoothing methods, i.e., BLF, WLS, TV, and  $L_0$  smoothing. Then, we apply ALM-Lp to cartoon artifacts removal and tongue image segmentation. The programs in our experiments are all coded in MATLAB and ran on a computer with Intel(R) Xeon(R) CPU E3-1230 V2@3.30GHz and 16GB memory.

### 4.1 ALM-Lp with Different $p$ Values

From Fig. 2, we can draw the similar conclusion with [7] that the gradient distributions of real-world images are well modeled by hyper-Laplacian prior, typically with  $0.5 \leq p \leq 0.8$ . Particularly, when  $p = 0.7$ , better tradeoff can be achieved in removing noise-like highlights and detailed textures and preserving salient edges. Thus we adopt  $p = 0.7$  for ALM-Lp in the following experiments.

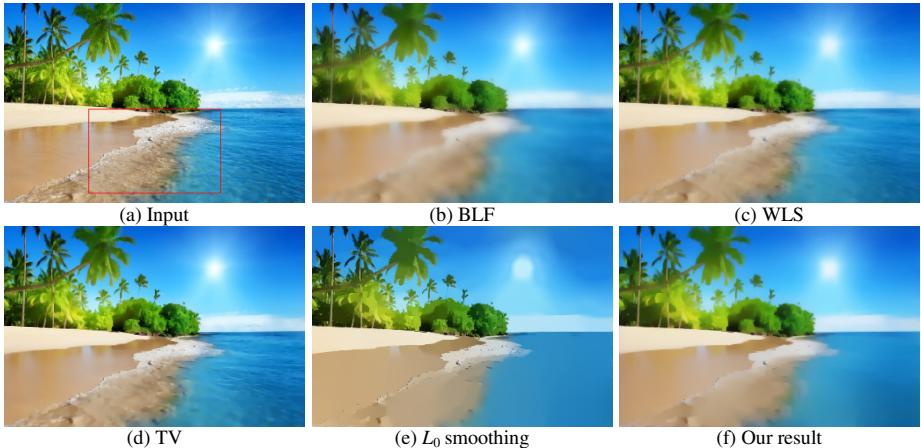


**Fig. 2.** The smoothing result of ALM-Lp with different  $p$  values. This figure is best viewed in electronic form and zoomed.

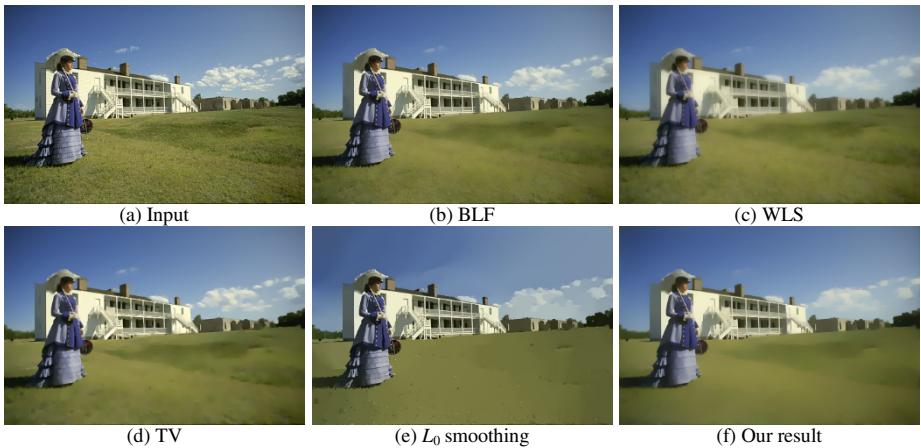
### 4.2 Comparison with Other Methods

Using six natural images, we compare ALM-Lp with several state-of the art image smoothing methods, i.e., bilateral filter, weighted least-squares, TV, and  $L_0$  smoothing. Fig. 3 and Fig. 4 show the results on images *beach* and *scenery*. On the whole, ALM-Lp can maintain the general structure and is effective in smoothing insignificant details, whether the gravel in Fig. 3 or the grass in Fig. 4.

We further compare the running time of ALM-Lp with that of  $L_0$  smoothing on the six natural images, as listed in Table 1. One can see ALM-Lp is more efficient than  $L_0$  smoothing.



**Fig. 3.** The smoothing results on the image *beach*. As the red boxes show, our method can better smoothing insignificant textures.



**Fig. 4.** The smoothing results on the image *scenery*

**Table 1.** Speed (sec.) comparison of  $L_0$  gradient smoothing and ALM-Lp method

Image	Size	$L_0$ smoothing	ALM-Lp
<i>rock</i>	$800 \times 533$	4.19	1.58
<i>flower</i>	$800 \times 533$	2.99	1.99
<i>beach</i>	$800 \times 533$	3.02	1.70
<i>pflower</i>	$475 \times 494$	1.45	0.97
<i>scenery</i>	$481 \times 321$	1.11	0.54
<i>basketball</i>	$270 \times 358$	0.72	0.31

### 4.3 Artifact Removal for Cartoon Pictures

Cartoon image compressed by conventional JPEG compression may contain some artifacts that severely damage the image structures. Thus, we need to remove artifacts while sharpening important salient edges. Fig. 5 shows the results obtained using ALM-Lp, from which we see the proposed method can effectively wipe off artifacts while preserving general structures, demonstrating its better property.

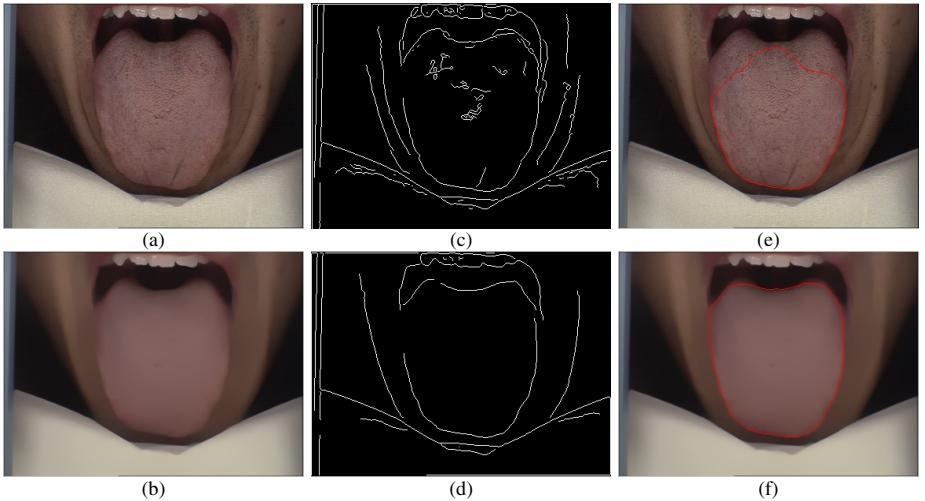


**Fig. 5.** Cartoon pictures artifacts removal

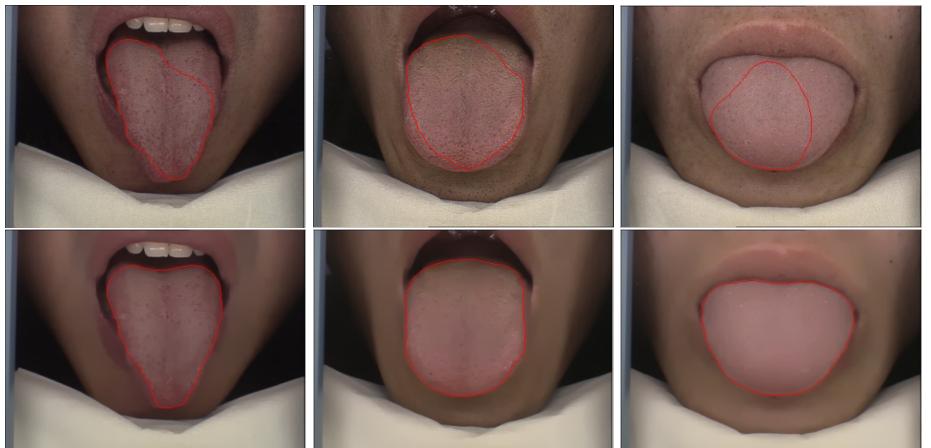
### 4.4 Tongue Images Segmentation

For tongue body segmentation, the edge maps obtained by edge detectors, e.g., Canny, often contain unnecessary textures, shown as Fig. 6(c), which makes it hard to correctly segment the tongue body. Thus, we first used ALM-Lp to enhance the edge of tongue body, whereas diminishing unnecessary details, and then the well-known gradient vector flow (GVF) snake [17] was adopted to perform the segmentation. From Fig. 6(c)(e), one can see that, wispy textures in the edge map of the original image make the snake curve converge to a wrong segmentation result, while the smoothed image using ALM-Lp has a remarkable improvement, in which unnecessary textures are diminished, leading to the satisfactory segmentation results shown as Fig. 6(d)(f).

Furthermore, we validate the segmentation performance on three more tongue images, as shown in Fig. 7, from which we can draw the conclusion that it makes a significant segmentation improvement to apply ALM-Lp as a preprocessing step.



**Fig. 6.** Tongue image segmentation. (a) the original tongue image, (b) the result using ALM-Lp, (c), (d) are edge maps, and (e), (f) are segmentation results, respectively.



**Fig. 7.** More tongue images segmentation results. The top line presents the results on original images, and the bottom line presents the results on smoothed ones obtained using ALM-Lp.

## 5 Conclusion

In this paper, we studied the image smoothing problem and proposed an ALM-L<sub>p</sub> method based on hyper-Laplacian gradient prior. An efficient augmented Lagrangian method (ALM) is developed to solve the proposed model. For natural image smoothing, ALM-L<sub>p</sub> can obtain better smoothing results while compared with the state-of-the-arts. That is to say, ALM-L<sub>p</sub> is effective in removing impulse noise-like highlights and detailed textures and preserving salient edges. Compared with  $L_0$  smoothing, ALM-L<sub>p</sub> is

more efficient. Moreover, ALM-Lp can also be applied to cartoon artifacts removal and tongue image segmentation. Since impulse noise-like reflection and highlights usually are inevitable in tongue images, ALM-Lp, as a preprocessing step, can significantly improve tongue image segmentation results.

## References

1. Park, J.M., Murphrey, Y.L.: Edge detection in grayscale, color, and range images. Wiley Encyclopedia of Computer Science and Engineering (2008)
2. Carreira, J., Sminchisescu, C.: Constrained parametric min-cuts for automatic objects segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 3241–3248 (2010)
3. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. ACM Transactions on Graphics (TOG) 21, 257–266 (2002)
4. Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. ACM Transactions on Graphics (TOG) 27, 67 (2008)
5. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1), 259–268 (1992)
6. Xu, L., Lu, C., Xu, Y., Jia, J.: Image smoothing via L0 gradient minimization. ACM Transactions on Graphics (TOG) 30(6), 174 (2011)
7. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-laplacian priors. NIPS 22, 1–9 (2009)
8. Field, D.J.: What is the goal of sensory coding? Neural Computation 6(4), 559–601 (1994)
9. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional-camera with a coded aperture. ACM Transactions on Graphics (TOG) 26(3), 70 (2007)
10. Simoncelli, E.P., Adelson, E.H.: Noise removal via bayesian wavelet coring. In: International Conference on Image Processing, vol. 1, pp. 379–382 (1996)
11. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In: IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3869–3872 (2008)
12. Cho, T.S., Joshi, N., Zitnick, C.L., Kang, S.B., Szeliski, R., Freeman, W.T.: A content-aware image prior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 169–176 (2010)
13. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted  $\ell_1$  minimization. Journal of Fourier Analysis and Applications 14(5-6), 877–905 (2008)
14. Zuo, W., Meng, D., Zhang, L., Feng, X., Zhang, D.: A generalized iterated shrinkage algorithm for non-convex sparse coding. In: IEEE International Conference on Computer Vision (ICCV) (2013)
15. She, Y.: An iterative algorithm for fitting non-convex penalized generalized linear models with grouped predictors. Computational Statistics & Data Analysis 56(10), 2976–2990 (2012)
16. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. NIPS 2, 6 (2011)
17. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. IEEE Trans. Image Processing 7(3), 359–369 (1998)

# Study on Distribution Coefficient in Regulation Services with Energy Storage System

Shaojie Tan\*, Xinran Li, Ming Wang, Yawei Huang, Tingting Xu  
and Xingting Cheng

College of Electrical and Information Engineering, Hunan University, Changsha, Hunan,  
China, 410082  
tansjhnu@163.com

**Abstract.** Renewable energies like wind power and PV come with great fluctuation and uncertainties, and their penetration into power system has brought great challenge to the security and stability of the grid. Energy storage system (ESS) possesses the ability to track power accurately with nearly no delay, which makes it capable of providing regulation services of high quality. Taking into account the State of Charge (SOC) of ESS, this paper analyzes the regulation signal and processes it with blocks of delay, filter, clipping and so on, presenting a method to determine the distribution coefficient of ESS and conventional generator (CG) dynamically based on the available regulation capacity, and an aggregative indicator is used to evaluate the relating effects and SOC maintaining effects. Simulation results show the method proposed can improve the regulation service and maintain the SOC within a desired range, providing technical support to the optimized operation of the grid.

**Keywords:** Regulation service, signal analyzing, signal processing, energy storage system, distribution coefficient.

## 1 Introduction

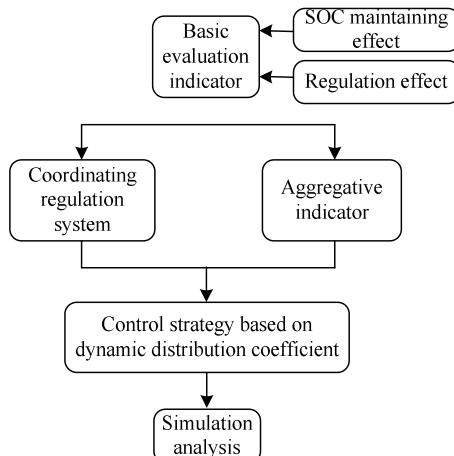
Renewable energies like wind power and PV come with great fluctuation and uncertainties, and their penetration into power system has brought great challenges to the security and stability of the grid [1-2]. Meanwhile, the slow reaction and low ramping rate of traditional generations fail to meet the needs of power system's fast development and absorb new energy. As a result, to safeguard frequency stability of power system in high permeability of renewable energy's generation is one of the new challenges for Chinese grid. In recent years, the rapid advancement of energy storage technologies has shed a new light on the solutions of these issues. Energy storage system (ESS) possesses the ability to track power accurately with nearly no delay [3-4], which makes it capable of providing regulation services of high quality. Being more efficient than conventional generations (CG), it may serve as a new solution to problems in the field of regulation services [5].

---

\* Corresponding author.

Published literature has done some research on control strategy of regulation service with energy storage system. Reference [6] analyzed the traditional ways of regulation services and the feasibility and advantages of electric vehicles (EV) participating in load frequency control, focused on the charging and discharging control of EVs. The method of proportional distribution control was used to control EVs to respond to the area control error and complete the frequency adjustment; reference [7] presented a fuzzy based frequency control strategy to distribute the regulation signal between PV systems, ESSs and EVs proportionally; references [8-9] proposed centralized Vehicle to Grid (V2G) control methods to participate in secondary frequency control. In summary, study on ESS providing regulation service is still at an exploratory stage. A systematic analysis of the distribution coefficients of ESS and CG has not been conducted and current studies only involve the static proportional distribution coefficients determination.

This paper analyzes the regulation in statistic ways and processes it with several blocks such as delay, filter, clipping and so on, presenting a method of determining the distribution coefficient of ESS and conventional generator based on the available regulation capacity, and an aggregative indicator is used to evaluate the relating effects and SOC maintaining effects. Results show the method proposed can effectively improve the regulation service and maintain the SOC within a desired range, providing technical support to the optimized operations of the grid. Fig.1 illustrates the framework of the study in this paper.

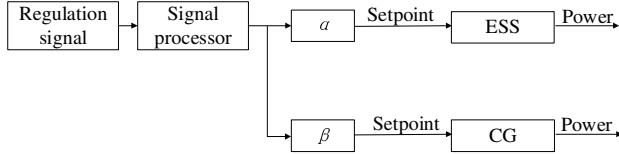


**Fig. 1.** Framework of the study

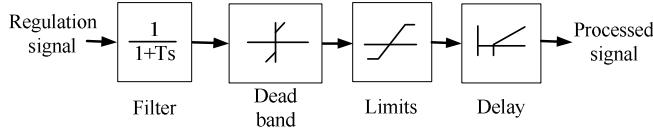
## 2 Distribution Coefficient

Owing to the imbalance of electricity supply and demand, the power grid needs to provide the appropriate regulation services, and the regulation signal is determined by the dispatch center. The analyzing and processing of regulation signal will be taken before sending it to ESS and CG, as is shown in Fig.2, where  $\alpha$  and  $\beta$  are distribution

coefficients of ESS and CG respectively. The signal processor with blocks like filter, dead band and so on is shown in Fig.3.



**Fig. 2.** Distribution coefficients of ESS and CG



**Fig. 3.** Signal processor

## 2.1 Traditional Determination of Distribution Coefficient

In the traditional way, distribution coefficients are determined according to the rated regulation capacity of ESS and CG proportionally

$$\begin{cases} \alpha = \frac{P_{\text{essrated}}}{P_{\text{essrated}} + P_{\text{cgated}}} \\ \beta = 1 - \alpha \end{cases} \quad (1)$$

where  $P_{\text{essrated}}$  and  $P_{\text{cgated}}$  are the rated regulation capacity of ESS and CG respectively.

## 2.2 Improved Determination of Dynamic Distribution Coefficient

In the improved way, distribution coefficients are determined according to the available (maximum) regulation capacity of ESS and CG dynamically

$$\begin{cases} \alpha = \frac{P_{\text{ess}}}{P_{\text{ess}} + P_{\text{cg}}} \\ \beta = 1 - \alpha \end{cases} \quad (2)$$

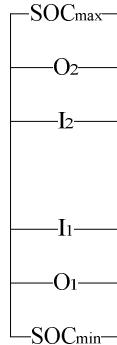
$$P_{\text{ess}} = \begin{cases} -P_{\text{essmin}}, & \text{Charge} \\ P_{\text{essmax}}, & \text{Discharge} \end{cases}$$

$$P_{\text{cg}} = \begin{cases} -P_{\text{cgmin}}, & \text{Charge} \\ P_{\text{cgmax}}, & \text{Discharge} \end{cases}$$

where  $P_{\text{ess}}$  and  $P_{\text{cg}}$  are maximum regulation capacity of ESS and CG respectively,  $P_{\text{essmax}}$  and  $P_{\text{cgmax}}$  are maximum regulation-up capacity of ESS and CG respectively,  $P_{\text{essmin}}$  and  $P_{\text{cgmin}}$  are maximum regulation-down capacity of ESS and CG respectively.

### Maximum Regulation Capacity of Energy Storage System

According to ESS's own characteristics, into five regions has been divided by O1, I1, I2 and O2 which can be valued at 0.3, 0.4, 0.6 and 0.7 respectively between the upper and lower limits of SOC, as is illustrated in Fig.4.



**Fig. 4.** Illustration of SOC region division

Different methods to determine distribution coefficients have carried out in various SOC region.

(1) Within the SOC band of [I1, I2]

If the SOC of ESS is within [I1, I2], the entire regulation signal is sent to ESS ( $\alpha=1$ ) unless it exceeds the rated regulation capacity of ESS when CG is needed to bear part of the regulation signal

$$\begin{cases} \alpha = \begin{cases} \frac{P_{\text{essrated}}}{P_{\text{reg}}}, & P_{\text{essrated}} \leq P_{\text{reg}} \\ 1, & P_{\text{essrated}} > P_{\text{reg}} \end{cases} \\ \beta = 1 - \alpha \end{cases} \quad (3)$$

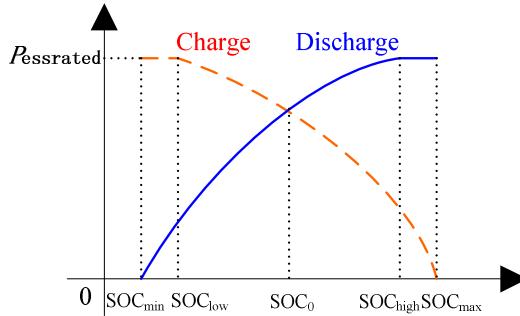
where,  $P_{\text{reg}}$  is the processed regulation signal.

(2) Within the SOC band of [O1, I1] or [I2, O2]

If the SOC of ESS is within [O1, I1] or [I2, O2], , an ESS SOC balance control considering the status of SOC is conducted to the determination of distribution coefficient and the ESS SOC balance control is shown in Fig.5.

$$\begin{cases} P_{\text{essmin}} = \begin{cases} P_{\text{essrated}} \left( 1 - \left( \frac{\text{SOC} - \text{SOC}_{\text{low}}}{\text{SOC}_{\text{max}} - \text{SOC}_{\text{low}}} \right)^n \right), & \text{SOC} \geq \text{SOC}_{\text{low}} \\ P_{\text{essrated}}, & \text{SOC} < \text{SOC}_{\text{low}} \end{cases} \\ P_{\text{essmax}} = \begin{cases} P_{\text{essrated}} \left( 1 - \left( \frac{\text{SOC} - \text{SOC}_{\text{high}}}{\text{SOC}_{\text{min}} - \text{SOC}_{\text{high}}} \right)^n \right), & \text{SOC} \leq \text{SOC}_{\text{high}} \\ P_{\text{essrated}}, & \text{SOC} > \text{SOC}_{\text{high}} \end{cases} \end{cases} \quad (4)$$

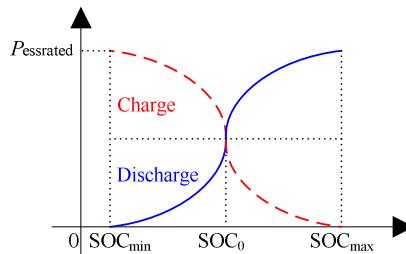
where  $SOC$  is state of charge of ESS;  $SOC_{max}$ ,  $SOC_{min}$ ,  $SOC_{high}$ ,  $SOC_{low}$  are upper limit, lower limit, higher value and lower value of ESS's SOC respectively, and they can be valued at 0.9, 0.1, 0.8 and 0.2 respectively;  $n$  is exponent of the equation, and it can be valued at 2.



**Fig. 5.** Relationship of maximum regulation capacity and SOC in SOC balance control

(3) Within the SOC band of  $[SOC_{min}, O1]$  or  $[O2, SOC_{max}]$

If the SOC of ESS is within  $[SOC_{min}, O1]$  or  $[O2, SOC_{max}]$ , an adaptive ESS SOC balance control considering the status of SOC more in-depth is conducted to the determination of distribution coefficient and the ESS SOC balance control is shown in Fig.6.



**Fig. 6.** Relationship of maximum regulation capacity and SOC in adaptive SOC balance control

If  $SOC_{min} < SOC \leq SOC_0$ , then

$$\begin{cases} P_{essmin} = \frac{1}{2} P_{essrated} \left( 1 + \sqrt{\frac{SOC - SOC_0}{SOC_{min} - SOC_0}} \right) \\ P_{essmax} = \frac{1}{2} P_{essrated} \left( 1 - \sqrt{\frac{SOC - SOC_0}{SOC_{min} - SOC_0}} \right) \end{cases} \quad (5)$$

If  $SOC_0 < SOC \leq SOC_{max}$ , then

$$\begin{cases} P_{essmin} = \frac{1}{2} P_{essrated} \left( 1 - \sqrt{\frac{SOC_i - SOC_0}{SOC_{max} - SOC_0}} \right) \\ P_{essmax} = \frac{1}{2} P_{essrated} \left( 1 + \sqrt{\frac{SOC_i - SOC_0}{SOC_{max} - SOC_0}} \right) \end{cases} \quad (6)$$

where  $SOC_0$  is the expected status of ESS's SOC, it is usually valued at 0.5.

### Maximum Regulation Capacity of Conventional Generator

In the determination of the dynamic distribution coefficient, the restriction on power ramp rate of CG should be considered

$$\begin{cases} P_{cg\min} = -\max \{-P_{cgated}, P_{cglast} - Ramp\Delta t\} \\ P_{cg\max} = \min \{P_{cgated}, P_{cglast} + Ramp\Delta t\} \end{cases} \quad (7)$$

where  $P_{cglast}$  is the previous power output of CG; Ramp is the ramp rate of CG,  $\Delta t$  is the interval of time.

## 3 Coordinating Regulation System

### 3.1 Energy Storage System Model

The ESS model includes blocks of regulation power, round trip efficiency, capacity limitation and energy calculation. After receiving the regulation signal, ESS responds to it immediately. Energy loss is existed owing the roundtrip efficiency. The model of ESS is shown in Fig.7.

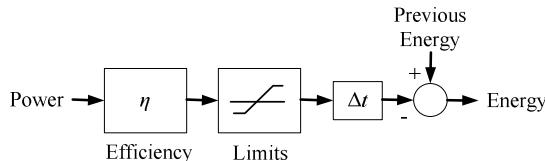
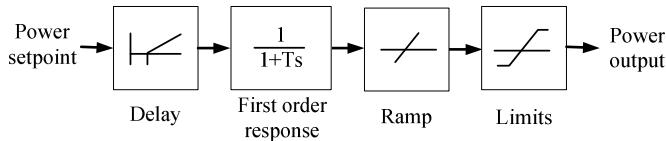


Fig. 7. ESS model

### 3.2 Conventional Generator Model

The CG model includes blocks of delay, first order response, power ramp rate limitation, regulation capacity limitation and power output. After receiving the regulation signal, CG responds to it slowly owing its delay, first order response and ramp rate. The model of CG is shown in Fig.8.



**Fig. 8.** CG model

## 4 Evaluation Indicator

### 4.1 Basic Indicator

(1) Indicator  $MAE_{rms}$  reflecting regulation effect

$$MAE_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^n (AE_i - AE_0)^2} \quad (8)$$

where  $AE_i$  is absolute error of each Sampling point;  $AE_0$  is the expected value of absolute error, it is usually valued at 0;  $n$  is the number of sampling points.

(2) Indicator  $SOC_{rms}$  reflecting SOC maintaining effect

$$SOC_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^n (SOC_i - SOC_0)^2} \quad (9)$$

where  $SOC_i$  is SOC status of each Sampling point;  $SOC_0$  is the expected value of absolute error, it is usually valued at 0.5;  $n$  is the number of sampling points.

### 4.2 Normalized Indicator

Due to the existence of different dimension inconsistency of indicators, the indicators should be normalized before evaluation.

(1) Normalized indicator  $MAE_{rms}'$  reflecting regulation effect

$$MAE_{rms}' = \frac{MAE_{rms} - MAE_{rmsmin}}{MAE_{rmsmax} - MAE_{rmsmin}} \quad (10)$$

where  $MAE_{rmsmax}$  and  $MAE_{rmsmin}$  are upper and lower limits of  $MAE_{rms}'$ .

(2) Normalized indicator  $SOC_{rms}'$  reflecting SOC maintaining effect

$$SOC_{rms}' = \frac{SOC_{rms} - SOC_{rmsmin}}{SOC_{rmsmax} - SOC_{rmsmin}} \quad (11)$$

where  $SOC_{rmsmax}$  and  $SOC_{rmsmin}$  are upper and lower limits of  $SOC_{rms}'$ .

### 4.3 Aggregative Indicator

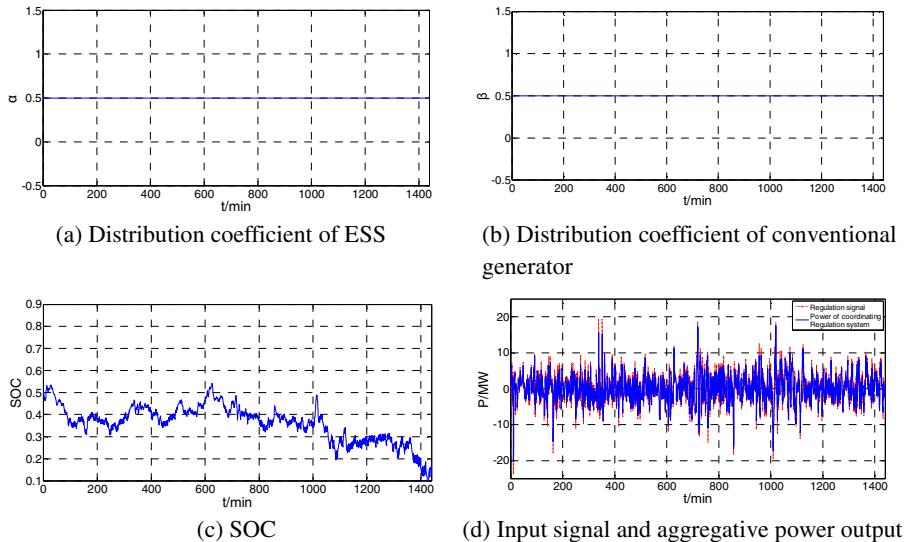
Considering regulation effect and SOC maintaining effect, and an aggregative evaluation indicator is conducted

$$K = k_1 \text{MAE}_{\text{rms}} + k_2 \text{SOC}_{\text{rms}} \quad (12)$$

where  $k_1$  and  $k_2$  are weight coefficients, both added up to 1, they can be valued at 0.5 respectively. The less K is, the better aggregative effect.

## 5 Case Study

In this case, actual area control error signals obtained from Hunan Electric Power Dispatching Control Center which has been processed by the signal processor are used as input signals of the coordinating regulation system with ESS and CG to compare the regulation effect and SOC maintaining effect of traditional static distribution coefficient and improved dynamic distribution coefficient. 1-minute signals of 1 day which were negated and normalized with the ratio of 15% for study have been used for the simulation study. In this paper, discharging power is positive while charging power is negative. Simulation results are shown in Fig.9 and Fig.10.

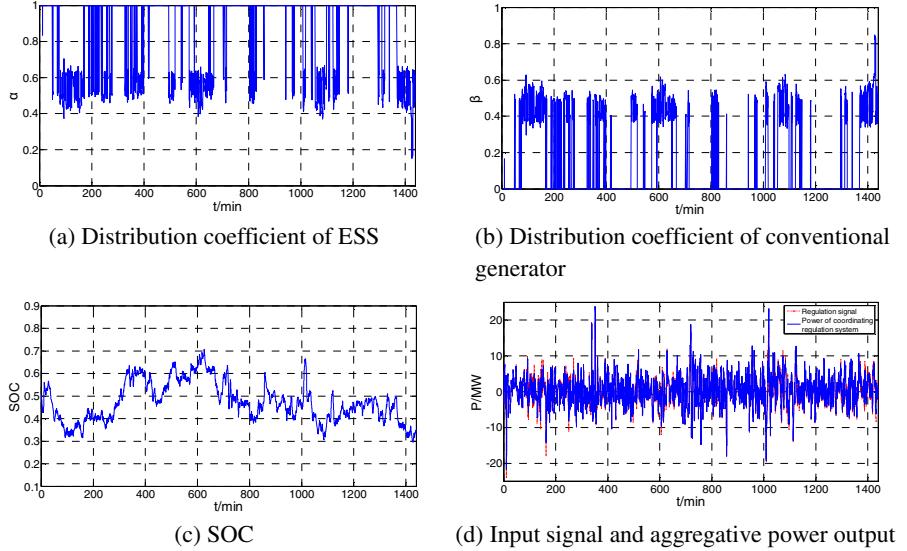


**Fig. 9. Traditional Static Distribution Coefficient**

Basic parameters have been set as follows:

ESS: Power range: -20MW~20MW; energy capacity: 5MWh; efficiency: 90%.

CG: Rated capacity: 400MW; regulation service range: -20MW~20MW; energy capacity: unlimited; ramp rate: 12MW/min (3% of rated capacity/min); delay time: 10s; first response time constant: 0.08s.



**Fig. 10.** Improved Dynamic Distribution Coefficient

The plots in Fig.9 and Fig.10 show the comparison between the traditional way and the improved way of distribution coefficient determination. The improved dynamic distribution coefficient method is able to adjust the distribution coefficient dynamically and maintain the ESS's SOC with 0.3~0.7 which is close to the expected value 0.5. Besides, from Fig.9 (d) and Fig.10 (d), the improved method can improve the regulation effects, tracking the regulation signal more accurately by shortening the deviation between regulation signal and power output of the coordinating system.

Using the evaluation indicator to assess the traditional method and the improved method, results are shown in Table 1.

**Table 1.** Evaluation results

Distribution coefficient	$MAE_{rms}$	$SOC_{rms}$	K
Traditional static distribution coefficient	0.2403	0.3990	0.3201
Improved dynamic distribution coefficient	0.1618	0.2287	0.1952

From Table 1, compared to the traditional way, the improved dynamic distribution coefficient method can maintain the SOC within a desired range and improve the regulation effect as well. The coordinating regulation system can provide regulation service with more stability and security by using the proposed method in this paper.

## 6 Conclusions

This paper analyzes the regulation in statistic ways and processes it with blocks of delay, filter, clipping and so on, proposing a method to determine the distribution coefficient of ESS and CG according to the available regulation capacity, and an aggregative indicator is used to evaluate the relating effects and SOC maintaining effects. Simulation results approve the correctness and validity of the method.

**Acknowledgement.** This work was financially supported by National High-tech Development and Research Plan (2011AA05A113) and National Key Fundamental Research Project (2012CB215100).

## References

1. Chang, J., Chen, D.: Idea of Energy Storage System participating in Frequency Regulation in National development. Study and Reference 46, 1–15 (2012) (in Chinese)
2. Wu, Z., Zeng, Y.G., Tie, H.: Wind/Solar Generation System and Energy Storage System. Chemical Industry Press, Beijing (2012) (in Chinese)
3. Chen, D., Zhang, L., Wang, S., et al.: Development of Energy Storage in Frequency Regulation Market of United States and Its Enlightenment. Automation of Electric Power System 37(1), 9–13 (2013) (in Chinese)
4. Eyer, J., Corey, G.: Energy Storage for the Electricity Grid: Benefits and Market Potential Assessment Guide, pp. 1–153. Sandia National Laboratories, Sandia (2010)
5. IEEE PES Committee Report: Current Operating Problems Associated with Automatic Generation Control. IEEE Transactions on Power Apparatus and Systems 98(1), 88–96 (1979)
6. Han, H.: The Study on the Control Strategy of V2G Participating Peak Regulation and Frequency Regulation of the Grid. Beijing Jiaotong University, Beijing (2011) (in Chinese)
7. Manoj, D., Tomonobu, S.: Fuzzy control of distributed PV inverters/energy storage systems/electric vehicles for frequency regulation in a large power system. IEEE Transactions on Smart Grid 4(1), 479–488 (2013)
8. Galus, M.D., Koch, S., Andersson, G.: Provision of load frequency control by PHEVs, controllable loads, a cogeneration unit. IEEE Transactions on Industrial Electronics 58(10), 4568–4582 (2011)
9. Ayakrishnan, R., Pillai, B.B.J.: Integration of Vehicle-to-Grid in the Western Danish Power System. IEEE Transactions On Sustainable Energy 2(1), 12–19 (2011)
10. Jin, C., Lu, N., Lu, S., et al.: A coordinating algorithm for distributioning regulation services between slow and fast power regulating resources. IEEE Transactions on Smart Grid (99), 1–8 (2013)
11. Jin, C., Lu, N., Lu, S., et al.: Coordinated control algorithm for hybrid energy storage systems. In: Proc. IEEE Power Energy Soc. Gen. Meet., pp. 1–7 (2011)

# All-Focused Light Field Image Rendering

Rumin Zhang<sup>1,\*</sup>, Yu Ruan<sup>2</sup>, Dijun Liu<sup>2</sup>, and Zhang Youguang<sup>1</sup>

<sup>1</sup>School of Electronic and Information Engineering, Beihang University, Beijing, China  
rm\_zhang@ee.buaa.edu.cn

<sup>2</sup>State Key Laboratory of Wireless Mobile Communications (CATT), Beijing, China  
Ruanyu814@163.com

**Abstract.** The coupling between aperture size and the depth of field (DOF) in traditional imaging remains one of the fundamental limits on photographic freedom. The emergence of the plenoptic camera imaging solves the issues. Based on the focus plenoptic camera implemented by Georgiev, we propose an all-focused image rendering algorithm which depends on the DOF. Using the raw image captured by the prototype of the camera, we successfully reconstruct an all-focused image of the scene and calculate a higher resolution depth map. Finally, we conclude that a large depth of field image can still be achieved even with a large aperture.

**Keywords:** Light field imaging, plenoptic camera, depth of field, full focus.

## 1 Introduction

Conventional camera does not record most of the information about the light impinging on the image sensor. It takes a sharp picture of the object positioned on the focal plane. But the image taken when the object is off the focal plane will be somewhat blurred. The coupling between aperture size and DOF (the range of depths that appears sharp in the resulting image) remains one of the fundamental limits on photographic freedom. On the one hand, a narrow aperture extends the depth of field and reduces blur of objects away from the focal plane. However, a narrow aperture not only requires a longer exposure which increases the blur due to the natural shake of our hands while the camera and the movement in the scene but also leads to a low SNR. On the other hand, in order to shorten the exposure time, wide aperture is required which reduce the DOF. How should we choose the correct aperture size?

The proposer of the plenoptic camera [1] (the light field camera, constructed with a main lens and internal micro-lens arrays fixed before the photo-sensor) solves the problem. The device can capture the lost information: to measure the full 4D light field including 2D spatial and 2D directional information and record the amount of light travelling along each ray that intersects at the sensor.

The use of micro-lens arrays in imaging is a technique called integral photography that was pioneered by Lippmann [2] and greatly refined by Ives [3,4]. The micro-lens

---

\* Corresponding author.

can produce tiny images focused on the main lens of the camera and the camera was named plenoptic camera first proposed by Adelson and Wang [5] which was further improved and implemented in a hand-held digital camera by Ng [6]. More specifically, the camera captures radiance with an array of micro-lenses. Each micro-lens samples a dense set of ray directions at a single spatial point. However, the limited resolution which is related to the number of micro-lens has been a significant drawback to the traditional plenoptic camera.

Georgiev and Lumsdaine[7] advanced the camera and presented a focused plenoptic camera. As the name implies, the focused plenoptic camera can produce images of much higher resolution than the traditional plenoptic camera which is treated as a relay system where the main lens creates a main image in the camera, and the image is remapped to the photo-sensor by the micro-lens array. Depending on the position of the main image in the camera, the focused plenoptic camera has two different modes of operation [8]: Keplerian and Galilean. And which mode is selected depends on object length.

As far, there are two rendering algorithms based on the plenoptic camera. The initial algorithm was introduced by Ng in his dissertation for the degree of Doctor of Philosophy [9]. In Ng's algorithm, the final image is calculated by summing the radiance function in all direction on each pixel. However, one micro-lens produces one pixel in the final image. If the micro-lens array in the plenoptic camera contains  $296 \times 296$  micro-lens, the resolution of the final image is just  $296 \times 296$ . Based on the focused plenoptic camera, Georgiev proposed a rendering algorithm depending on the DOF. In the rendered image, different pixel patches (not the pixel) selected based on the DOF.

This paper is organized as follows: Section II presents the structure of the plenoptic camera and its imaging principle. We estimate the depth and derive the rendering algorithm in Section III. In section IV and V, the DOF analysis and the experiment on a raw image captured by the camera are completed. The last section concludes the paper.

## 2 Plenoptic Camera

As the Fig.1 denotes, the focused plenoptic camera consists of main lens, micro-lens and photo-sensor.  $cz$  represents the distance between the main lens and micro-lens array.  $b$  is the distance from micro-lens to photo-sensor.  $a$  and  $z'$  represent the distance from the main lens image plane to micro-lens array and to the main lens, respectively.  $F$  is the focal length of the main lens.

In the focused plenoptic camera, the main lens maps the 3D world of the scene into a 3D world inside of the camera. The mapping is governed by the lens equation. Assuming  $z$  is the object length, we can obtain:

$$\frac{1}{z} + \frac{1}{z'} = \frac{1}{F} \quad (1)$$

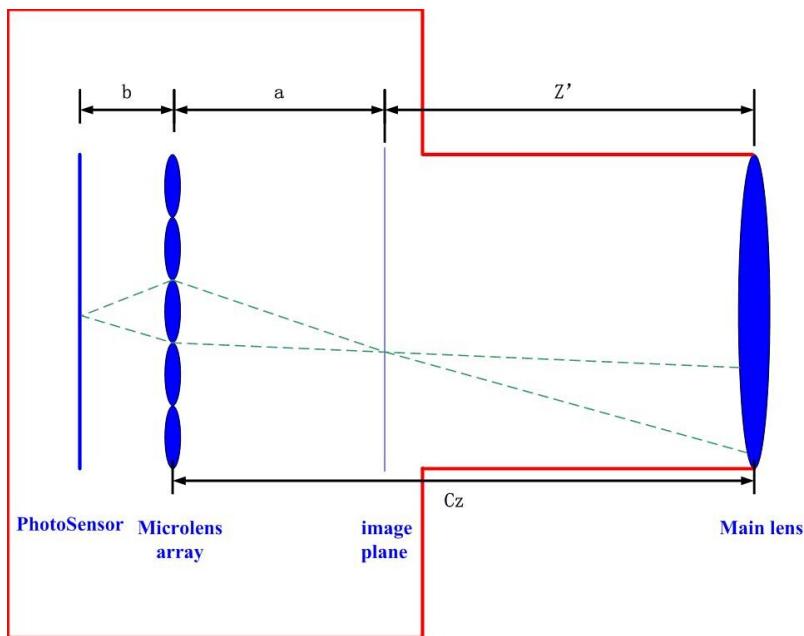
Where  $z$  is the distance from the main lens to the object plane, then

$$z' = \frac{Fz}{z - F} \quad (2)$$

In the focused plenoptic camera, as the  $c_z$  is a constant for a camera, when  $z' < c_z$ , the image plane will be pointed in front of the micro-lens array. Thus, the object length can be denoted as

$$z > \frac{Fc_z}{c_z - F} \quad (3)$$

Where,  $F=80\text{mm}$ ,  $c_z=100\text{mm}$ , thus  $z>400\text{mm}$ , the image plane would always locates before the micro-lens array. In the conventional photography, object length is always greater than 1m. All the cases discussed in this paper are based on the object length larger than 400mm.



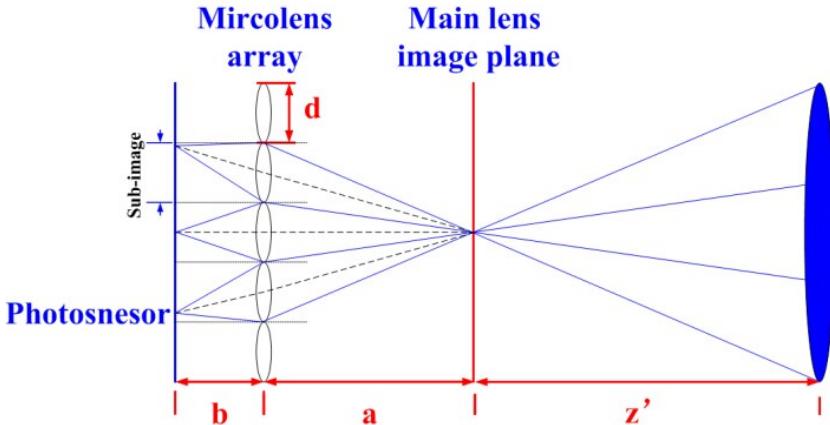
**Fig. 1.** The focused plenoptic camera model [8]

## 2.1 The Imaging Principle of the Camera

As the Fig.2 depicts, the micro-lens width is  $d$ . Let  $N$  represent the number of the micro-lens, then the photo-sensor can be divided into  $N$  patches. The image captured

by each patch is a sub-image of the final image.  $r_n(i_n)$  is the radiance function of the sub-image, where  $0 < n < N, 0 < i_n < m$  and  $m$  is the number of the pixels corresponding to each micro-lens and  $m = d/\delta$  ( $\delta$ -the size of each pixel).

In addition,  $a_n$  describes the object length corresponding to the  $n$ -th micro-lens. And the position of the sub-image depends on  $a_n$ . As a result, in order to render the all-focused image, calculating the DOF of each sub-image is critical in this paper.



**Fig. 2.** The principle of the focused plenoptic camera [10]

### 3 Depth Estimation and Image Rendering

#### 3.1 Depth Estimation

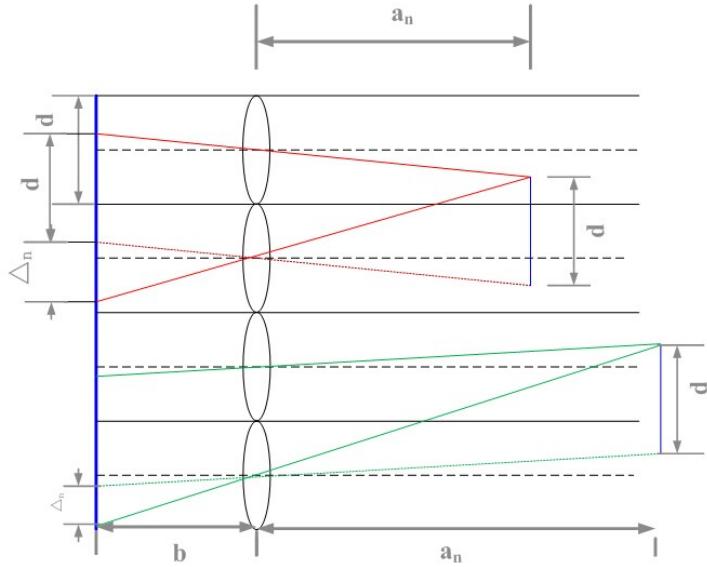
The depth information required in this section describes the distance between the image plane and the micro-lens array which can be calculated through the disparity between the adjacent sub-images. According to a recent taxonomy, stereo algorithm that generates depth measurements can be roughly divided into two classes, namely global and local algorithms. Global algorithms on the basis of the minimization of a global const function carry out disparity assignments, while local algorithms also referred to as area-based algorithm, calculate the disparity at each pixel on the basis of the photometric properties of the neighboring pixels.. In this paper, SAD(sum of Absolute Difference) algorithm is employed.

Based on the SAD, the disparity between the sub-images can be obtained.

$$\sigma(\Delta_n) = \frac{1}{M} \sum_{i_n=1}^{i_n=m'-\Delta_n} |r_n(i_n + \Delta_n) - r_{n+1}(i_n)| \quad (4)$$

Here,  $\Delta_n$  is the disparity of pixels between adjacent sub-images.  $M$  is the number of the matching pixels.  $m'$  represents pixels number of one sub-image.

Disparity of the sub-image can be found by minimizing the above equation through changing the  $\Delta n$ . As the Fig.3 illustrates, the depth  $a$  could be calculated with the optimal  $\Delta n$ .



**Fig. 3.** The disparity of the object between the neighbor views formed by micro-lens

According to the principle of similar triangle, we can get

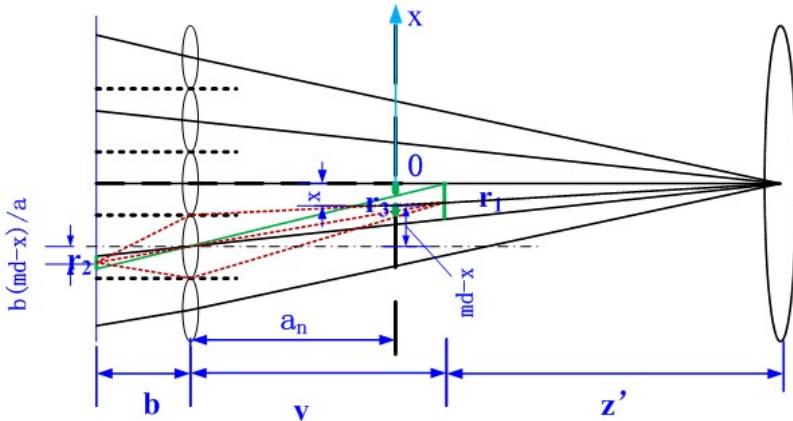
$$\frac{d}{\Delta_n \delta} = \frac{a_n}{b} \quad (5)$$

$$a_n = \frac{bd}{\Delta_n \delta} = \frac{m}{\Delta_n} b \quad (6)$$

### 3.2 Rendering Algorithm

Based on the assumption of Lambertian, all the lights coming from one point have an equal radiance. Assuming an arbitrary point Z on the object scene, as Fig.4 demonstrates, all the lights emitting from the point converge at a point on the conjugate image plane a distance v in front of micro-lens array.

To show how the focused plenoptic camera captures radiance, we derive the expressions for the image at the sensor and the rendering plane in terms of the radiance. Let  $r_1(x)$  be the radiance at the conjugate image plane and  $r_2(x)$  and  $r_3(x)$  be the radiance at sensor and at the rendering plane, respectively.



**Fig. 4.** Geometric diagram

Assuming the center of the coordinate locates at the center of the micro-lens array, we can derive the following formulas,

$$r_2 \left( \frac{b(md-x)}{a_n} \right) = r_1(x) \quad (7)$$

$$r_3 \left( \left( 1 + \frac{a_n - v}{z'} \right) x \right) = r_1(x) \quad (8)$$

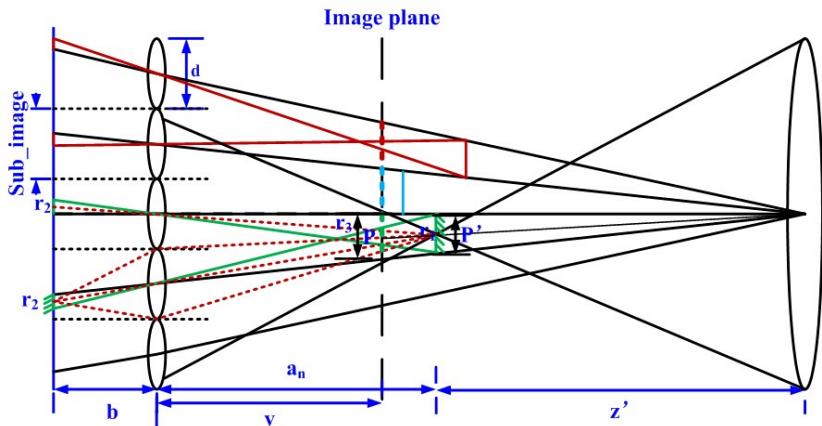
$$\text{Here, } -\frac{N}{2}d < x < \frac{N}{2}d$$

What we obtained are the continuous radiance functions. However, the sample of the photo-sensor is discrete. In next section, we will deduce the discrete radiance function.

As the Fig.5 shows,  $p$  and  $p'$  are the widths of the images on the rendering plane and on the conjugate image plane corresponding to the  $n$  sub-image on the sensor, respectively. Using the geometric information,  $p$  and  $p'$  can be calculated.

$$p = \frac{z' + a_n - v}{z' + a_n} d \quad (9)$$

$$p' = \frac{z'}{z' + a_n} d \quad (10)$$



**Fig. 5.** Rendering algorithm for reconstructing all-in-focus image

Furthermore, the width of the sub-image on the sensor (the hatched part in green) can be derived.

$$p'' = \frac{b}{a_n} \frac{z'}{z' + a_n} d \quad (11)$$

$\delta$  is the diameter of the pixel and  $p''_n$  is the number of the pixels on the sensor covered by the n sub-image.

$$p''_n = \frac{b}{a_n} \frac{z'}{z' + a_n} \frac{d}{\delta} \quad (12)$$

The width of the sub-image corresponding to each micro-lens is d and the sub-image covers m pixels. Then the relation between the n sub-image( on the final rendering plane) and the n micro-lens image can be expressed as

$$r_2(i_n) = r_n(i_n - \frac{d}{2\delta} + \frac{nd}{\delta} + c_n) \quad (13)$$

Here,  $0 < i_n < p_m''$ ,  $c_n$  represents the correction parameter.

$$c_n = \left( m - \frac{N+1}{2} \right) \frac{b}{z' + a_n} \frac{d}{\delta} \quad (14)$$

Furthermore, the image on the conjugate image plane of the main lens can be deduced

$$r_1(i_n) = r_2(p''_m - i_n + 1), 0 < i_n \leq p''_m \quad (15)$$

Finally, the final output image is calculated

$$r_3(i_n) = r_2(-i_n + p''_m - \frac{d}{2\delta} + \frac{nd}{\delta} + c_n + 1) \quad (16)$$

Where  $0 < i_n \leq p''_m$ .

The all-focused rendering algorithm can be summarized as following:

1. Read the data from the photo-sensor and get the:

$$r_2(i_n), 0 < n < N, 0 < i_n < \frac{d}{\delta}$$

2. Estimate the depth information of the image on the conjugate image plane:

$$a_n, 0 < n < N$$

3. Use  $a_n$  and calculate the image on the image plane

$$r_1(i_n) = r_2(p''_m - i_n + 1) \quad 0 < i_n < p''_n$$

4. Calculate the size of the above image:

$$p = \frac{z' + a_n - v}{z' + a_n} \cdot d$$

By interpolation, we get a new image:

$$r_1'(x_{i_n}), 0 < i_n < n, 0 < x_{i_n} < \frac{p}{\delta}$$

5. Calculate the final image:

$$r_3(i_n) = r_3(i_n \cdot \frac{d}{p_n}) = r_2' \left( \frac{z'}{c} \cdot \frac{d}{p'_n} \cdot i_n \right)$$

$$0 \leq n < N, 0 \leq i_n < p'_n$$

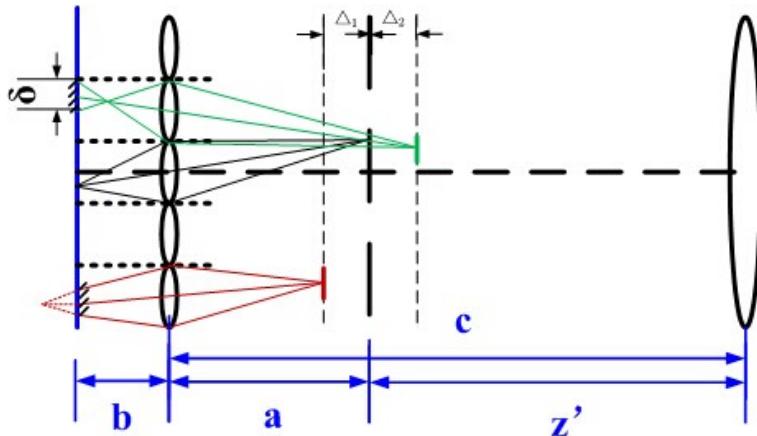
## 4 DOF Analysis

As we have mentioned,  $\delta$  is the pixel diameter and  $f$  is the focal length of the micro-lens.  $c$  represents the distance from main lens to the micro-lens array. In order to avoid overlap between the adjacent pixels, the circle of confusion on the sensor should be restricted within the diameter of pixel  $\delta$ . As shown in Fig.6, the DOF of the micro-lens is calculated.

$$\Delta_1 = \frac{a^2\delta}{bd+a\delta} \quad \Delta_2 = \frac{a^2\delta}{bd-a\delta} \quad (17)$$

So the range of  $z'$  is expressed as

$$c-a-\Delta_1 < z' < c-a+\Delta_2 \quad (18)$$



**Fig. 6.** The DOF of the micro-lens

Using the thin lens formula, we can deduce the DOF of the main lens.

$$\frac{F(c-a+\Delta_2)}{c-a+\Delta_2-F} < z < \frac{F(c-a-\Delta_1)}{c-a-\Delta_1-F} \quad (19)$$

Here we adopt a set of parameters in the conventional camera.  $F=80\text{mm}$ ,  $A=F/2.8$ ,  $\delta=7.2\mu\text{m}$   $b=1.3\text{mm}$ ,  $c=97.5\text{mm}$ ,  $f=1.2\text{mm}$ ,  $d=0.36\text{mm}$

$$\frac{1}{b} + \frac{1}{a} = \frac{1}{f} \quad (20)$$

$$a = \frac{bf}{b-f} = 15.6\text{mm} \quad (21)$$

Replacing the unknown parameters in (17) with the above constants,  $\Delta_1 = 3\text{mm}$ ,  $\Delta_2 = 4.77\text{mm}$ , referring to (19).

$$1.04m < z < \infty \quad (22)$$

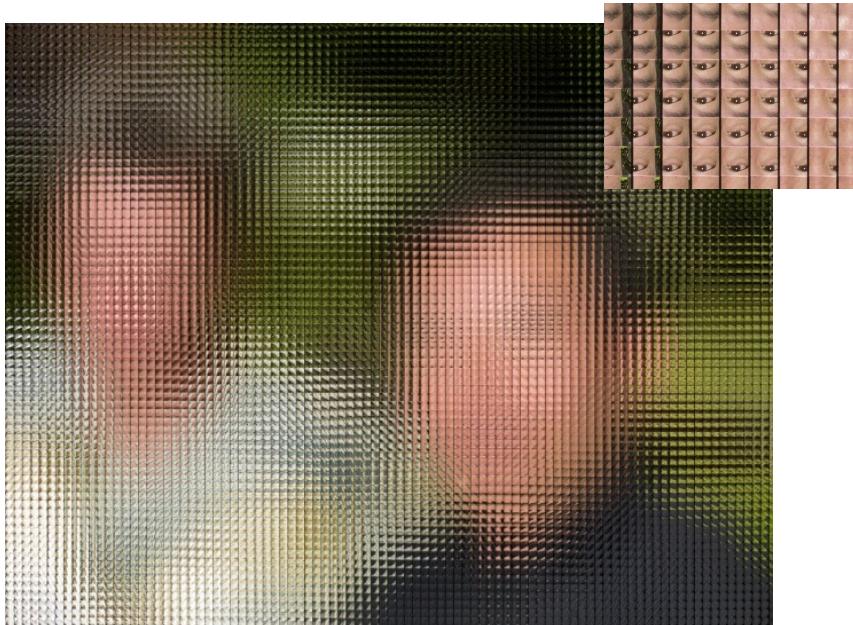
As for a conventional camera,  $F = 80\text{mm}$   $A = F / 2.8$   $\delta = 7.2\mu\text{m}$   $z = 5\text{m}$ , similarly, we get  $\Delta_1' = 80\text{mm}$ ,  $\Delta_2 = 77\text{mm}$ , furthermore:

$$4.92m < z < 5.08 \quad (23)$$

## 5 Simulation Result

As shown in Fig.7, (a) is the raw image captured by the focused plenoptic camera which is proposed by T.Georgiev. The raw image consists of  $96 \times 72$  sub-images. From the enlarged region shown at the corner, we can find that there is disparity between the adjacent sub-images. With a large DOF, the image mapping on the photo-sensor would focus better.

Based on the all-focused rendering algorithm proposed in this paper, we render a full-focused image as the shown in Fig.7(b). Fig.7 (c) is the depth map of the image which is calculated by the algorithm described in the section 3.1.

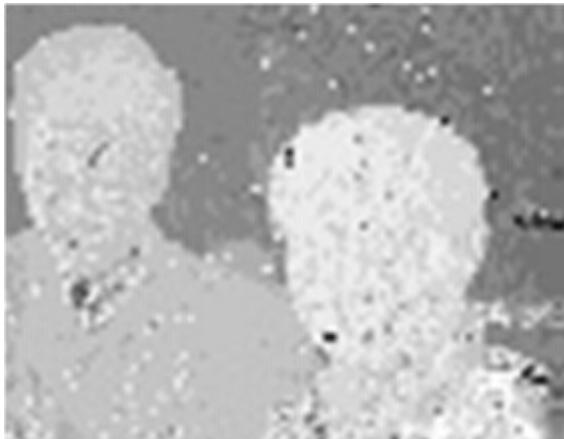


(a) The raw image from capture by the focused plenoptic camera

**Fig. 7.** The raw image of the plenoptic camera and the simulation



(b) All-focused image rendered by the algorithm



(c) Depth map

**Fig. 7.** (Continued)

## 6 Conclusion

In this paper we have presented the analysis of the focused plenoptic camera structure and calculate the DOF of the camera. With the disparity between the adjacent sub-images, we can render an all-focused image using a raw image captured by the camera. The focused plenoptic camera is absolutely much more practical.

**Acknowledgments.** This work was supported by State Key Laboratory of wireless communication. The author wish to thank T.Georgiev for sharing his original image captured by the focused plenoptic camera, from which we verified our algorithm.

## References

1. Gortler, S., et al.: The Lumigraph. In: Proc. of ACM SIGGRAPH, pp. 43–54 (1996)
2. Lippmann, G.L.: Photographie Integrale. Comptes Academie des Sciences 146, 446–551 (1908)
3. Ives, H.E.: Parallax Panoramagrams made with a large diameter lens. Opt. Soc. Amer. 20, 332–342 (1930)
4. Ives, H.E.: Optical properties of a Lippmann lenticulated sheet. J. Opt. Soc. Amer. 21, 171–176
5. Adelson, T., Wang: Single lens stereo with a plenoptic camera. IEEE Trans. Pattern Anal. Machine Intell. 14, 99–106 (1992)
6. Ng, L.M., et al.: Light field photography with a hand-held plenoptic camera. In: CTSR02. Stanford University Computer Science, California (2005)
7. Lumdsdaine, A., Georgiev, T.: Full Resolution Light-field Rendering. Teach.rep, Adobe Systems (January 2008)
8. Lumdsdaine, A., Georgiev, T.: The focused plenoptic camera. In: Proc. Int. Conf. on Computational Photography, Stanford University, pp. 1–11 (2009)
9. Ng, R.: Digital Light Field Photography. Stanford University, California (2006)
10. Georgiev, T., Lumdsdaine, A.: Depth of Field in Plenoptic Cameras. In: EUROGRAPHICS (2009)

# Hyperspectral Image Unmixing Based on Sparse and Minimum Volume Constrained Nonnegative Matrix Factorization

Denggang Li, Shutao Li, and Huali Li

College of Electrical and Information Engineering,  
Hunan University, Changsha, 410082, China  
`{lidenggang, shutao_li, lihuali}@hnu.edu.cn`

**Abstract.** Hyperspectral Unmixing (HU) aims at getting the endmember signature and their corresponding abundance maps from highly mixed Hyperspectral image. Nonnegative Matrix Factorization (NMF) is a widely used method for HU recently. Traditional NMF only take sparse constraint or minimum volume constraint into consideration leading to unmixing results not accurately enough. In this paper, we propose a new method based on NMF through combining volume constraint with sparse constraint. According to the convex geometry, we impose minimum volume constraint on endmember matrix. Because sparsity is nature property of abundance, we add the sparse constraint on abundance matrix. Both the experiments on synthetic and real scene images show the effectiveness of the proposed method.

**Keywords:** Hyperspectral unmixing, nonnegative matrix factorization, minimum volume constraint, sparse constraint.

## 1 Introduction

With the rapid development of satellite and remote sensing technology, hyperspectral image has received more and more attention for its numerous spectral bands which can provide more information for spectral analysis. Traditional remote sensing technology can be used to agriculture supervision, geographical exploration, military application, etc. However, due to the limit of spatial resolution, mixed pixels are wildly existed in hyperspectral images. Mixed pixels means a single pixel contains several materials. The existences of mixed pixels influence our study of spectrums seriously. Therefore, the goal of hyperspectral unmixing (HU) is to decompose the mixed pixels into each material—called endmembers and their corresponding fraction abundance [1][2][3].

Recently, many algorithms have been proposed to solve the HU problems mentioned above. Those algorithms usually can be divided into two categories. One is called Two Step Methods (TSM) and the other is called Single Step Methods (SSM). TSM firstly extract the endmember from the mixed pixels, and then estimate the corresponding abundances. Common endmember extraction algorithms include VCA [4], N-FINDR [5], PPI [6], etc. After getting the endmember we usually use FCLS [7] to obtain the abundance maps. SSM can get the endmembers and abundance maps

simultaneously. Traditional SSM include ICA [8], MVCNMF [9], SMCNMF [10], GLNMF [11], NMFSSC [12], etc.

Compared with the two unmixng methods mentioned above we can see that, for TSM, if there existing error in the first step of endmemer extraction then the final abundance results will become worse. So, SSM will avoid this problem. Moreover, in real hyperspectral image the assumption of pure pixels existed is not always hold true. That is to say the above mentioned methods with assumption of pure pixels existed for unmimixng such as VCA, N-FINDER, PPI, etc, can't always extract endmembers successfully. To overcome these difficulties, blind spectral unmixing was proposed. NMF was originally proposed for data dimension reduction [13] and has been widely used for hyperspectral unmixing recently. MVCNMF [9] exploits the convex geometry information of mixed pixels and adds volume constraint on NMF. However, this method neglects the sparsity of the abundance. Sparsity means that mixed pixel only contained a few of endmembers and sparsity is a nature property of hyperspectral images. NMFSSC [12] incorporated the sparse and smooth constraints to NMF. Nevertheless, this method neglects the convex geometry information of mixed pixels. We were taking advantages of the MVCNMF and NMFSSC method, both volume constraint on endmembers, sparse constraint on abundance were combined in this paper.

The rest of this paper is organized as follows. In section 2, we introduced the Hyperspectral unmixing and NMF algorithm model. In section 3, original MVCNMF and NMFSSC have been discussed. Then we have proposed the new sparse and minimum volume constrained NMF. Both synthetic image and real hyperspectral image experiments have proven the improvement of our methods than original method in Section 4. Finally, conclusions have been conducted in Section 5.

## 2 Hyperspectral Unmixing and NMF

In the real hyperspectral scene there usually existed two models: linear mixed model (LMM) and nonlinear mixed model (NLMM).To put it simply, LMM means mixed process happens at a macroscopic level [14], and photons only interact with one material before reaching the satellite sensor [15]. Although NLMM is widely existed in real scene, for simplicity, we only discussed LMM in this paper.

### 2.1 LMM

LMM plays an important role in the fields of hyperspectral unmixing. In the following, we give the mathematic model of LMM:

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \quad (1)$$

where  $\mathbf{X} \in \mathbf{R}^{n \times m}$  denotes the observed hyperspectral with n spectral bands and m pixels. Matrix  $\mathbf{A} \in \mathbf{R}^{n \times l}$  represents endmemer signatures and  $\mathbf{S} \in \mathbf{R}^{l \times m}$  is the corresponding fraction abundance. l are the number of endmembers. Finally  $\mathbf{N}$  denote the possible noises and errors. Blind spectral unmixing means both the  $\mathbf{A}$ ,  $\mathbf{S}$ ,  $\mathbf{N}$  are unknown.

There are some constraints in the real hyperspectral scene. 1) Because abundance maps pixel cannot be negative, so the elements of matrix of  $\mathbf{S}$  must be nonnegative:  $s_{ij} \geq 0, i = 1 \dots l, j = 1 \dots m$ . 2) The sum of abundance satisfies sum-to-one constraint:

$$\sum_{i=1}^l s_{ij} = \mathbf{1}.$$

## 2.2 NMF

Nonnegative matrix factorization (NMF) was proposed by Lee, D. and Seung, S. [13] in 1999. Then NMF was applied to dimension reduction, face recognition and document cluttering, etc. Given a nonnegative matrix  $\mathbf{Y} \in \mathbf{R}^{n \times m}$  and  $r < \min(n, m)$ , the goal of NMF is to find two nonnegative matrices  $\mathbf{W} \in \mathbf{R}^{n \times r}$  and  $\mathbf{H} \in \mathbf{R}^{r \times m}$  approximate fulfill the following equation:

$$\mathbf{Y} \approx \mathbf{WH} \quad (2)$$

Then we use the Euclidean distance as our objective function to minimize the error between  $\mathbf{Y}$  and  $\mathbf{WH}$

$$\min f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{WH}\|^2 \quad (3)$$

Compared NMF and LMM we can see that NMF model can be used for unmixing. So, adds the nonnegative and sum-to-one constrained on equation (3), objective function  $f$  subject to.  $\mathbf{W} \succeq 0, \mathbf{H} \succeq 0, \mathbf{1}_r^T \mathbf{H} = \mathbf{1}$ , where  $\succeq$  denote elementwise inequality.

## 3 NMF with Sparse and Minimum Volume Constraints

The solution of the object function proposed in equation (3) is not unique and this model hasn't rigorous physical meaning. If  $\mathbf{W}$  and  $\mathbf{H}$  are solutions of objective function and we can finding a reversible matrix  $\mathbf{D}$  satisfying  $\mathbf{1}_r^T \mathbf{D}^{-1} \mathbf{H} = \mathbf{1}$ , then  $\mathbf{WD}$  and  $\mathbf{D}^{-1} \mathbf{H}$  are solutions too. To solve this problem many physical meaning constrained NMF has been proposed: minimum volume constrained NMF (MVCNMF) [9], minimum distance constrained NMF(MDCNMF) [16], Manifold Regularized Sparse NMF [11], NMF base on sparse and smooth constraints NMFSSC [12],  $L_{1/2}$  sparsity NMF [17] etc.

MVCNMF [9] is a classical constrained NMF method that put constraint on end-member matrix. This method was to finding the minimum volume of simplex that formed by mixed data cloud. The objective function was revised as

$$\begin{aligned} \min f(\mathbf{A}, \mathbf{S}) &= \frac{1}{2} \|\mathbf{X} - \mathbf{AS}\|^2 + \text{vol}(\mathbf{J}(\mathbf{A})) \\ \text{s.t. } \mathbf{A} &\succeq 0, \mathbf{S} \succeq 0, \mathbf{1}_c^T \mathbf{S} = \mathbf{1}. \end{aligned} \quad (4)$$

where  $\text{vol}(\mathbf{J}(\mathbf{A}))$  represent the volume of simplex determined by endmember matrix. MVCNMF gives a meaningful constraint on endmember matrix, but this method neglect the sparsity of abundance, which is a nature property of abundance.

NMFSSC [12] is a new constrained method for spectral unmixing, this method exploits sparseness and smoothness property of abundance. The NMFSSC model is as follows:

$$\begin{aligned} \min f(\mathbf{A}, \mathbf{S}) &= \|\mathbf{X} - \mathbf{AS}\|^2 - \alpha_1 J_s^1(\mathbf{S}) + \alpha_2 J_s^2(\mathbf{S}) \\ &= \|\mathbf{X} - \mathbf{AS}\|^2 - \alpha_1 \sum_{i=1}^r \sum_{j=1}^n f(s_{ij}) + \alpha_2 \|\mathbf{DS}\|_2 \\ \text{s.t. } \mathbf{A} &\succeq 0, \mathbf{S} \succeq 0, \mathbf{1}_c^T \mathbf{S} = \mathbf{1}. \end{aligned} \quad (5)$$

where  $f(s_{ij}) = \theta(s_{ij} - \frac{1}{r})$ ,  $\mathbf{D}$  meaning distributional gradient,  $J_s^1(\mathbf{S})$  and  $J_s^2(\mathbf{S})$  represent sparse and smooth constraint respectively. Experiment results in [12] show that puts sparse and smooth constraints on abundance matrix can lead to slightly better results than original NMF. However, this method neglects the geometry information of endmembers, which is an important factor for spectral unmixing.

In this paper, we combine the MVCNMF mentioned in [9] with sparse constrained NMF in [12] together and proposed our new method called sparse and minimum volume constrained NMF (SMVCNMF). SMVCNMF model is given as follows:

$$\begin{aligned} \min f(\mathbf{A}, \mathbf{S}) &= \frac{1}{2} \|\mathbf{X} - \mathbf{AS}\|^2 + \alpha_1 J_A(\mathbf{A}) - \alpha_2 J_s(\mathbf{S}) \\ \text{s.t. } \mathbf{A} &\succeq 0, \mathbf{S} \succeq 0, \mathbf{1}_r^T \mathbf{S} = \mathbf{1} \end{aligned} \quad (6)$$

where  $J_A(\mathbf{A}) = \frac{\tau}{2} \det^2(\mathbf{C} + \mathbf{BU}^T(\mathbf{A} - \boldsymbol{\mu}\mathbf{1}_c^T))$  represents the volume of simplex.  $\tau$  is a constant,  $\boldsymbol{\mu}$  is data mean of  $\mathbf{X}$ ,  $\mathbf{C} = \begin{bmatrix} \mathbf{1}_r^T \\ \mathbf{0} \end{bmatrix}$ ,  $\mathbf{B} = \begin{bmatrix} \mathbf{0}_{r-1}^T \\ \mathbf{I} \end{bmatrix}$ ,  $\mathbf{U} \in \mathbf{R}^{n \times (r-1)}$  are  $r-1$  principal components from  $\mathbf{X}$  through PCA.  $J_s(\mathbf{S}) = \sum_{i=1}^r \sum_{j=1}^m \theta(s_{ij} - \frac{1}{r})^2$  represents the sparse constraint.  $\theta$  is constant too.

After proposing the objective function, then we give the optimization algorithm of our method. The first problem is how to initializing the endmember matrix  $\mathbf{A}$  and abundance matrix  $\mathbf{S}$ , here we resort to VCA [4] and FCLS [7] for our initialization. Then the next problem is to computing matrix  $\mathbf{A}$  and  $\mathbf{S}$ . Similarly to [9], [10], [11], [12] we use alternating update rules, that is to say fixed one matrix while alternating updating the other matrix:

$$\begin{aligned} \mathbf{A}^{k+1} &= \arg \min_{\mathbf{A}} f(\mathbf{A}, \mathbf{S}^k) \leq f(\mathbf{A}^k, \mathbf{S}^k) \\ \mathbf{S}^{k+1} &= \arg \min_{\mathbf{S}} f(\mathbf{A}^{k+1}, \mathbf{S}) \leq f(\mathbf{A}^{k+1}, \mathbf{S}^k) \end{aligned} \quad (7)$$

Use the gradient descent updating rules [18], then above equations can be changed into:

$$\begin{aligned}\mathbf{A}^{k+1} &= \max(0, \mathbf{A}^k - \alpha^k \nabla_{\mathbf{A}} f(\mathbf{A}^k, \mathbf{S}^k)) \\ \mathbf{S}^{k+1} &= \max(0, \mathbf{S}^k - \beta^k \nabla_{\mathbf{S}} f(\bar{\mathbf{A}}^{k+1}, \mathbf{S}^k))\end{aligned}\quad (8)$$

where function  $\max(0, X)$  used to ensure nonnegative constraint and  $\bar{\mathbf{A}}^{k+1} = \begin{bmatrix} \mathbf{A}^{k+1} \\ \mathbf{1}_r^T \end{bmatrix}$

used to ensure sum-to-one constraint.  $\alpha^k$  and  $\beta^k$  are steepsizes selected according to Armijo rules [19]. The gradient about matrices  $\mathbf{A}$  and  $\mathbf{S}$  are  $\nabla_{\mathbf{A}} f(\mathbf{A}, \mathbf{S}) = (\mathbf{AS} - \mathbf{X})\mathbf{S}^T + \alpha_1 \tau \det^2(\mathbf{Z})\mathbf{UB}^T(\mathbf{Z}^{-1})^T$ ,  $\mathbf{Z} = \mathbf{C} + \mathbf{BU}^T(\mathbf{A} - \mu \mathbf{1}_c^T)$ .  $\nabla_{\mathbf{S}} f(\mathbf{A}, \mathbf{S}) = \mathbf{A}^T(\mathbf{AS} - \mathbf{X}) - 2\alpha_2 \theta(\mathbf{S} - \frac{1}{r})$ . Finally, use maximum iteration number and tolerated error as stopping rules.

SMVCNMF algorithm was summarized as follows:

---

**Algorithm 1.** Sparse and Minimum Volume Constrained NMF (SMVCNMF)

---

**Input:** nonnegative matrix  $\mathbf{X}$  and endmembers numbers  $r$ .

**Output:** endmember matrix  $\mathbf{A}$  and abundance matrix  $\mathbf{S}$

**Step 1:** initialize matrix  $\mathbf{A}$  and  $\mathbf{S}$  and set parameters  $\alpha_1, \alpha_2, \theta, \tau$ .

**Step 2:** normalize  $\mathbf{S}$  sum-to-one constraint and augment  $\mathbf{A}$  to  $\bar{\mathbf{A}}$ .

**Step 3:** while stopping rules do not meet updating matrix  $\mathbf{A}$  and  $\mathbf{S}$  according to equation (8).

**Step 4:** while stopping rules satisfied, stop updating and obtained the final results.

---

## 4 Experiments

In this section, both synthetic and real hyperspectral image experiments have been conducted to test the performance of the proposed approach. Firstly, we propose performance metrics to evaluating algorithms. Spectral angle distance (SAD) is a widely used metric [9],[10],[11],[12],[16],[17] to compare performance of diffident methods. The definition of SAD is:

$$\text{SAD} = \frac{180}{\pi} \cos^{-1} \left( \frac{\mathbf{a}^T \hat{\mathbf{a}}}{\|\mathbf{a}\|_2 \|\hat{\mathbf{a}}\|_2} \right) \quad (9)$$

Where  $\mathbf{a}$  and  $\hat{\mathbf{a}}$  represent true and estimated endmemer respectively. The smaller value of SAD that the better results.

### 4.1 Synthetic Image

To generate the synthetic image, we random choose 4 endmembers from USGS [20] library. Every spectrum has 224 bands and wavelengths range from 0.38 to  $2.5 \mu\text{m}$ .

The synthetic image with  $64 \times 64$  pixels was divided into  $8 \times 8$  small blocks and every small block have pure endmember signature. Then use a  $7 \times 7$  filter to generate synthetic mixed data, replace all pixels whose abundance is larger than 0.8 with 4 endmembers, and their abundance equals to 1/4, to ensure that there is no pure pixels existing. Finally added the noise to synthetic image for simulate the possible errors, here we add zero-mean Gaussian noise with SNR=20 dB. In our methods we set parameter  $\alpha_1 = 1, \alpha_2 = 0.06, \theta = 0.1, \tau = 0.08$ , the number of maximum iterations is 150, tolerated error is  $10^{-6}$ . Table 1 has compared the SAD between VCA, MVCNMF, L<sub>1/2</sub>NMF and SMVCNMF. Because the smaller SAD means the better results, from the table we can see that the VCA algorithm performances the worst among those methods. This is because VCA algorithm was to find the pure endmembers from hyperspectral images, however, this assumption is not true in our experiment. So VCA can't obtain endmembers correctly. On the contrary, NMF is blind unmixing methods and doesn't need the assumption of pure pixel existed. The performances of NMF are better than VCA. When compared with SMVCNMF, L<sub>1/2</sub>-NMF and MVCNMF, we can find that our new method is better than other methods, mainly due to other methods neglect sparsity of abundance or volume constraint of endmembers.

**Table 1.** SAD between VCA, MVCNMF, L<sub>1/2</sub>-NMF and SMVCNMF in synthetic experiment

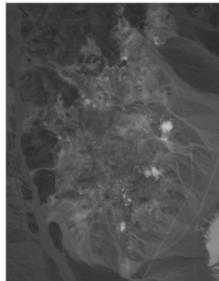
	VCA	MVCNMF	L <sub>1/2</sub> NMF	SMVCNMF
Endmember1	3.7753	0.3238	1.0554	<b>0.3063</b>
Endmember2	6.2046	1.0421	1.4800	<b>0.9024</b>
Endmember3	13.6902	1.3085	2.9065	<b>1.1703</b>
Endmember4	6.5448	0.5698	1.2914	<b>0.4937</b>
Mean	7.5537	0.8110	1.6833	<b>0.7182</b>

## 4.2 Real Hyperspectral Image

In this section, we use real hyperspectral image collected by AVIRIS over Cuprite [20] to test the proposed algorithm. The original image consists of 250 lines, 191 pixels per lines, and contained 224 bands, after removing the water-vapor absorbed bands only 188 bands left for our experiment. Fig 1 shows the 30th bands of real hyperspectral image. The number of endmembers was chose to 9 according to virtual dimension (VD). Use VCA and FCLS as our starting points for matrices  $\mathbf{A}$  and  $\mathbf{S}$ .

A quantitative comparison of SAD between VCA, MVCNMF, L<sub>1/2</sub>NMF and the proposed method has been listed on Table 2. Table 2 shows NMF having better results than VCA in most cases. In real hyperspectral scenes, the existence of pure pixels is not always true, so VCA algorithm can't always extracted endmembers correctly and result is the worst. Instead, NMF needn't the assumption of pure pixels existed, the results of NMF are better than VCA. From Table 2 we also can see that our new method SMVCNMF has good performance than other single constrained NMF methods (MVCNMF, L<sub>1/2</sub>NMF) when comparing mean value of SAD. This is mainly due to MVCNMF ignoring the sparsity of abundance, and L<sub>1/2</sub>NMF neglects convex geometry information of endmembers. Adding sparse constraint and volume constraint on NMF can lead to better

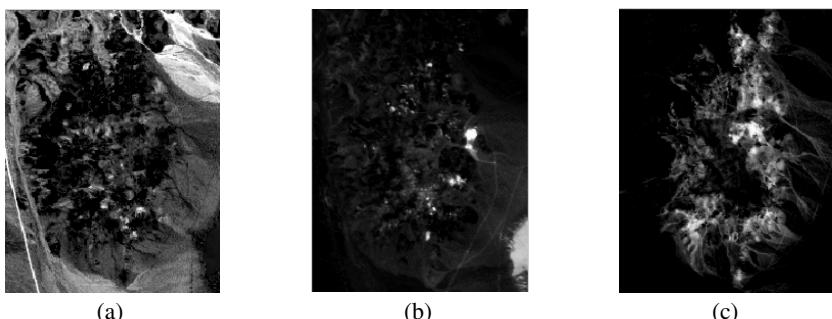
results compared to single constrained NMF. What's more, the estimated abundance maps of our new method SMVCNMF were illustrated in Fig.2. Compared with published results [21], we can clearly see that our abundance maps are more closely to the published results than other methods.



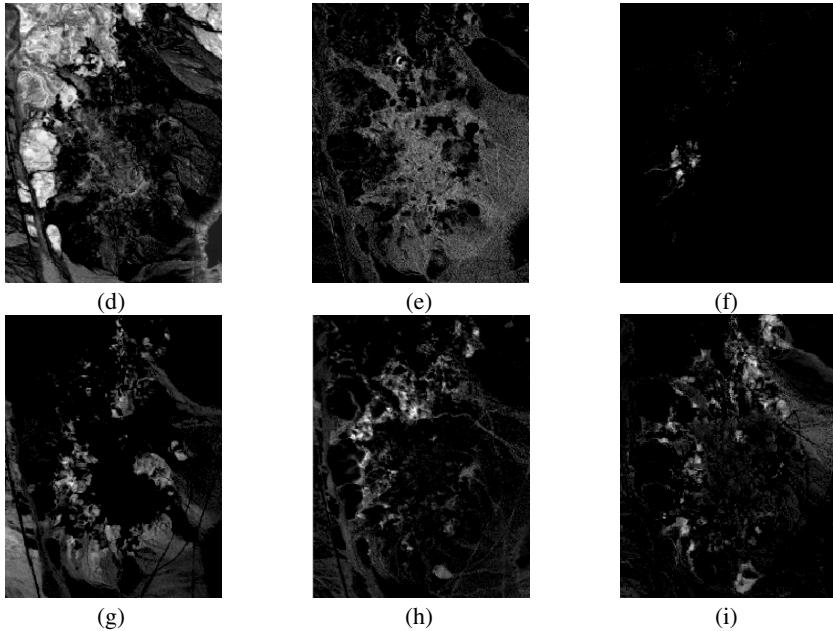
**Fig. 1.** Band 30 of Cuprite data image

**Table 2.** SAD between VCA, MVCNMF, L<sub>1/2</sub>-NMF and SMVCNMF in real hyperspectral experiment

	VCA	MVCNMF	L <sub>1/2</sub> -NMF	SMVCNMF
Alunite	22.0805	<b>15.2863</b>	17.2658	15.6919
Chalcedony	<b>8.0403</b>	9.0823	11.4980	8.5894
Muscovite	7.9701	<b>6.3739</b>	7.3755	6.4523
Desert varnish	10.4291	6.3370	<b>5.3062</b>	5.8007
Montmorillonite	<b>11.7112</b>	12.1258	13.4674	12.9237
Buddingtonite	8.3417	6.0264	7.3098	<b>5.8857</b>
Nontronite#1	7.5282	8.7829	12.4330	<b>6.9107</b>
Nontronite#2	7.8546	7.9196	9.1392	<b>7.3993</b>
Nontronite#3	10.0863	<b>6.4764</b>	6.8359	6.7240
Mean	10.5101	8.7123	10.0701	<b>8.4864</b>



**Fig. 2.** The abundance maps using SMVCNMF. (a) alunite; (b) chalcedony; (c) muscovite; (d) desert\_varnish; (e) montnorillonite; (f) buddingtonite; (g) nontronite#1; (h) nontronite#2; (i) nontronite#3.



**Fig. 2. (Continued)**

## 5 Conclusions

In this paper, an improved method sparse and minimum volume constrained NMF algorithm (SMVCNMF) was proposed for spectral unmixing. Here we utilize the properties of endmembers and abundance, then combined sparse constraint with volume constraint to NMF algorithm. Both synthetic and real hyperspectral image experiments have shown the improvement of our new method compared with original MVCNMF method.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant No. 61172161 and No. 61301255, the National Natural Science Foundation for Distinguished Young Scholars of China under Grant No. 61325007.

## References

1. Keshava, N.: A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal* 14(1), 55–78 (2003)
2. Zhu, F., Wang, Y., Xiang, S., Fan, B., Pan, C.: Structured sparse method for hyperspectral unmixing. *ISPRS Journal of Photogrammetry and Remote Sensing* 88, 101–118 (2014)

3. Dias, J.M.B., et al.: Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(2), 354–379 (2012)
4. Nascimento, J.M.P., Dias, J.M.B.: Vertex component analysis: a fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* 43(4), 898–910 (2005)
5. Winter, M.E.: N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. In: Proc. of the SPIE, vol. 3753, pp. 266–275 (1999)
6. Boardman, J.W., Kruse, F.A., Green, R.O.: Mapping target signatures via partial unmixing of AVIRIS data. In: Proc. Summaries JPL Airborne Earth Sci. Workshop, pp. 23–26 (1995)
7. Heinz, D.C., Chang, C.-I.: Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* 39(3), 529–545 (2001)
8. Wang, J., Chang, C.-I.: Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* 44(9), 2601–2616 (2006)
9. Miao, L., Qi, H.: Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Trans. Geosci. Remote Sens.* 45(3), 765–777 (2007)
10. Yang, Z., Zhou, G., et al.: Blind spectral unmixing based on sparse nonnegative matrix factorization. *IEEE Trans. on Image Processing* 20(4), 1112–1125 (2011)
11. Lu, X., Wu, H.: Manifold regularized sparse NMF for hyperspectral unmixing. *IEEE Trans. Geosci. Remote Sens.* 51(5), 2815–2826 (2011)
12. Wu, C., Shen, C.: Spectral unmixing using sparse and smooth nonnegative matrix factorization. In: Proc of 21st International Conference on Geoinformatics (GEOINFORMATICS), pp. 1–5 (2013)
13. Lee, D., Seung, S.: Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755), 788–791 (1999)
14. Singer, R.B., McCord, T.B.: Mars: Large scale mixing of bright and dark surface materials and implications for analysis of spectral reflectance. In: Proc. 10th Lunar and Planetary Science Conf., pp. 1835–1848 (1979)
15. Clark, R.N., Roush, T.L.: Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res.* 89(7), 6329–6340 (1984)
16. Yu, Y., Guo, S., Sun, W.: Minimum distance constrained nonnegative matrix factorization for the endmember extraction of hyperspectral images. In: Proc. of SPIE, vol. 6790, p. 679015 (2007)
17. Qian, Y., Jia, S., Zhou, J., Robles-Kelly, A.: Hyperspectral unmixing via L1/2 sparsity-constrained nonnegative matrix factorization. *IEEE Trans. Geosci. Remote Sens.* 49(11), 4282–4297 (2011)
18. Lin, C.-J.: Projected Gradient Methods for Nonnegative Matrix Factorization. MIT Press Neural Computation 19(10), 2756–2779 (2007)
19. Bertsekas, M.P.: Constrained optimization and lagrange multiplier methods. Academic, New York (1982)
20. <http://aviris.jpl.nasa.gov/>
21. [http://speclab.cr.usgs.gov/PAPERS/tetracorder/FIGURES/fig9b\\_cuprite95.gif.2.2um\\_map.gif](http://speclab.cr.usgs.gov/PAPERS/tetracorder/FIGURES/fig9b_cuprite95.gif.2.2um_map.gif)

# An Adaptive Harris Corner Detection Algorithm for Image Mosaic

Haixia Pan, Yanxiang Zhang, Chunlong Li, and Huafeng Wang

Colleage of Software, Beihang University, Beijing, China

{haixiapan, wanghuafeng}@buaa.edu.cn,  
stdcoutzyx@163.com, lcl426@gmail.com

**Abstract.** Image Stitching refers to the technology fusing more than one images with overlapping part into a large field of view image. Image mosaic consists of image preprocessing, image registration and image fusion. To solve problems of serious clustering phenomenon and fewer corner points in the texture region caused by traditional Harris Corner detection algorithm, this paper proposes an improving adaptive threshold setting algorithm by calculating the second-order value of the corner response function, avoiding effects of the selection of scale factor  $k$  and threshold  $T$  on corner detection. To overcome the weakness of obvious traces in the jointing places caused by traditional weighted average method for image fusion, this paper enhances the weighted average method with trigonometric functions. Experimental results show our proposed algorithms can effectively eliminate the gap generated by image mosaic, with a better speed and precision.

**Keywords:** Image Stitching, Image Registration, Corner Detection, Image Fusion.

## 1 Introduction

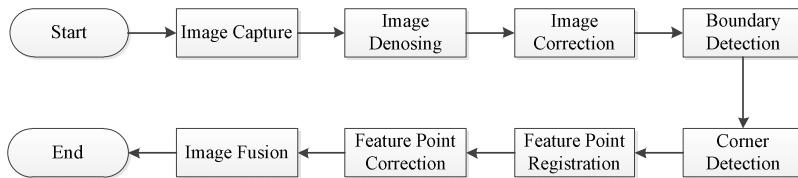
Image Stitching [1] refers to the technology fusing more than one images with overlapping part into a large field of view image. Nowadays, with the popularization of the intelligent equipment, the high definition and wide-angle image is becoming increasingly urgent. Image Stitching technology, known as one of the newest achievements in the image processing, has gained lots of attentions from researchers and is developing at a high speed recent years. Image Stitching technology has played an important role in aeronautics, astronautics, geological exploration, video session, medicine and military, and also stands as a hotspot in computer visual analogue, computer effects and augmented reality research [2,3,4,5,6].

Image mosaic consists of image preprocessing, image registration and image fusion [7]. As the base of image mosaic is the preprocessing, cores are the registration and fusion, deciding the precision, speed and visual quality. The image mosaic is divided into registration based on feature and registration based on gray level, depending on the image information used in registration. The registration based on feature [8] shows stronger adaptability in gray level transformation, deformation and

exposure discrepancy, which additionally can locate the matching positions easily and accurately. Although with highly accuracy, the one on gray level, which is also named as correlation matching algorithm, is hard to meet the demand of instantaneity because of its large computation and complexity [9]. The normal features of the registration are points, lines, close-contoured and other advanced ones such as Gaussian Sphere [10]. More attentions are attracted by registration based on feature points on account of its characteristics as being easily for fetching and less sensitive from image deformation. In the field of registration on feature points, the Harris Corner Detection algorithm [11] is famous for its high stability and robustness, but it has failed to consider the impact of the selection for the scale factor  $k$  and the threshold  $T$  of the algorithm. This paper proposes an improved Harris algorithm for this. While the traditional weighted average method is intuitionistic, fast, and less sensitive from discrepancy of exposure, shutter phenomenon [12] appears if there exists objects in the overlap region, we enhances weighted average method with trigonometric functions for image fusion, effectively reducing the obvious traces in the image splicing place.

## 2 Pipeline for Image Mosaic

Image Mosaic consists of image preprocessing, image registration and image fusion. For the possibility that there will be some effects from image rotation, image scaling and disparity of exposure, images are needed to be preprocessed before registration, whose keys are speed and precision. Finally images are stitched into one, in which no obvious traces are allowed for a good result. Figure 1 shows how the process flows.



**Fig. 1.** Flow chart for image mosaic

### 2.1 Image Preprocessing

Because of the flaw on image obtaining equipment and noise from outer condition, images are interfered during the digitization and transmission, resulting in the noisy output. Median filter is often used to denoise the image. As a nonlinear smoothing technique, it can help maintain the boundary information effectively. The principle of median filter is to replace the value of specified point with the mid-value of its neighborhoods, making the surrounding pixels close to the real value, thus eliminating the isolated noise points. Median filter is calculated as formula 1.

$$g(x, y) = \text{med}\{f(x - k, y - l)\}, (k, l \in W) \quad (1)$$

In formula 1,  $f(x, y)$ ,  $g(x, y)$  stand for the original image and the processed respectively.  $W$  is the two-dimension template, usually regions of  $2 \times 2$ ,  $3 \times 3$  are

used, any other different shapes like threadiness, roundness, cross, annulus can also serve the same purpose.

During the generating process, because of the nonlinearity of imaging system, or change of the shooting visual angle, the images will have discrepancy with each other, mainly on the gray level and geometrical deformation. Normalization [9] is able to make the amendment. After that most discrepancy of the gray level between images can be eliminated, thus decreasing the deviation in image registration.

## 2.2 Feature Point Extraction

As one of the most important parts in image mosaic, the precision of image registration affects the image fusion directly. Although the one based on gray level is more intuitionistic and easy to implement, it focuses on the specified template of image for stitching, so the result will be interfered by the smooth cover caused by the similar measurement. Additionally, the registration is also affected by the scaling transformation, rotation, overlapping, especially illumination. Registration deviation emerges since the nonlinear inhomogeneous illumination. Therefore, the registration based on gray level isn't widely used in image mosaic. However, the image registration based on feature points can overcome the shortcomings of the one on gray level, making it widely used. Extraction of image feature greatly speeds up with reducing the amount of calculation and can maintain the image during the image displacement, scaling, rotation, and other transformations. Properties of unique and easily distinguish of the image features make it easy to detect the positions' changes, thus highly increasing the precision of registration. This paper mainly focuses on the image registration based on feature points. In this section traditional Harris Corner Detection Algorithm is discussed, and then comes the algorithm mentioned in this paper. Experiments and analysis of those results are given out in the next section.

### Traditional Harris Corner Detection Algorithm

Harris Corner Detection [10] algorithm was proposed in 1988 by C. Harris and M. J. Stephens. The brightness variation  $E(\mu, v)$  is the local autocorrelation function of a pixel  $I(x, y)$ 's local offset  $(\mu, v)$ , as showed in formula 2.

$$E(\mu, v) = \sum_{x,y} w(x, y)[I(x + \mu, y + v) - I(x, y)]^2 \quad (2)$$

In formula 2, the  $I(x, y)$  stands for the gray level function.  $[I(x + \mu, y + v) - I(x, y)]$  refers to the gradient value of gray level.  $w(x, y)$  is the window function, which represents the weight of each pixel. There are two different window functions, namely two-valued and Gaussian. The influence of the noise from rectangular window can be reduced by Gaussian function, whose two-dimension version is given.

$$w(x, y) = \exp [-(x^2 + y^2)/2\sigma^2] \quad (3)$$

When the partial excursion  $(\mu, v)$  is small, Second-order Taylor series expansion of  $I(x + \mu, y + v)$  on  $(x, y)$  can be made in formula 4 after taking out the leading term.

$$E(\mu, v) \cong [\mu \nu] M \begin{bmatrix} \mu \\ \nu \end{bmatrix} \quad (4)$$

$M$  is a  $2 \times 2$  symmetric matrix, also called autocorrelation matrix. It can be represented by formula 5, in which  $\otimes$  refers to the convolution.

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = e^{-\frac{(x^2+y^2)}{2\sigma^2}} \otimes \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} A & C \\ C & B \end{bmatrix} \quad (5)$$

$$X = I \otimes [-1 \ 0 \ -1] = \frac{\partial I}{\partial x} = I_x; \quad Y = I \otimes [-1 \ 0 \ -1]^T = \frac{\partial I}{\partial y} = I_y \quad (6)$$

In formula 6,  $I_x$  and  $I_y$  refer to the gradient on horizon and verticality. The extremum curvature of the gray level autocorrelation function for particular pixel can approximation be represented by the eigenvalue of matrix  $M$ . Supposing two eigenvalues are  $\lambda_1$  and  $\lambda_2$ . If both of them are large, namely both the extremum curvature of the horizontal and vertical autocorrelation function for the pixel are large, the pixel can be treated as a corner. On the other hand, if both are small, the pixel is in a planar region. If one is large while the other is small, the region is boundary. In practice, in order to avoid solving the eigenvalue directly to increase efficiency, the response function of corner can be defined as follows.

$$R = \det M - k(\text{trace}M)^2 \quad (7)$$

In formula 7,  $\det M = \lambda_1 \lambda_2 = AB - C^2$ ,  $\text{trace}M = \lambda_1 + \lambda_2 = A + B$ . The scale factor  $k$  depends on experience. In most condition, it is selected from 0.04 to 0.06. If  $R$  is larger than the default threshold  $T$ , the point can be judged a corner point.

Nobel has improved the function to avoid the deviation causing by the scale factor  $k$  [13]. He defined the function  $R$  in formula 8.

$$R = \frac{\det M}{(\text{trace}M)^2} = \frac{AB-C^2}{A+B} \quad (8)$$

### Improved Harris Corner Detection Algorithm

Although Harris operator is classic, there are still some deficiencies. For example, though the scale factor  $k$  is defined in a particular range, the result of feature detection is poor when large differences exist between images. After getting the partial extremum, comparison with the threshold  $T$  and corner function  $R$  need to be made to determine the corner points, during which it's hard to select the threshold  $T$ . Too small of  $T$  leads to too many feature points and cluster phenomenon while too large of  $T$  corresponds to too few feature points, which decreases the accuracy of the results. This paper puts forward an adaptive Harris Corner Detection algorithm for that.

Because of the difficulty for the selection of threshold  $T$ , an improvement based on the Noble algorithm is made in formula 9, in which  $M$  is the autocorrelation matrix.

$$R(\mu, \nu, \sigma_I, \sigma_D) = \frac{\det(M(\mu, \nu, \sigma_I, \sigma_D))}{\text{trace}(M(\mu, \nu, \sigma_I, \sigma_D))} \quad (9)$$

$$M(\mu, \nu, \sigma_I, \sigma_D) = \sigma_D^2 G(\sigma_I) \otimes \begin{bmatrix} L_\mu^2(\mu, \nu, \sigma_D) & L_\mu L_\nu(\mu, \nu, \sigma_D) \\ L_\mu L_\nu(\mu, \nu, \sigma_D) & L_\nu^2(\mu, \nu, \sigma_D) \end{bmatrix} \quad (10)$$

$$L(\mu, \nu, \sigma_D) = G(\sigma_I) \otimes I \quad (11)$$

In the formula above,  $\sigma_I$  is the integral scale factor,  $\sigma_D$  is the differential scale factor,  $G(\sigma)$  represents the Gaussian function with variance of  $\sigma$  and mean value of 0,  $L_u$  and  $L_v$  refer to the partial derivative of direction U and direction V, det and trace are the abbreviation for the determinant and trace of matrix.

Template is calculated by mean circulation filter with the corner function R, integral scale factor  $\sigma_I$  and the differential scale factor  $\sigma_D$ , which can be used for second-order statistic to get the max value Max and the great value tmpMax. The pixel should be the maximum coordinate  $I_{max}$  when the corner function R is equal to the max value Max and Max isn't equal to the tmpMax. The threshold T can be calculated as formula 12, in which R is from formula 9.

$$T = R \cdot \max(I_{max}) \quad (12)$$

The steps for the algorithm are as follows.

1. Calculate the integral scale  $\sigma_I$  and the differential scale  $\sigma_D$ ;
2. Figure out the partial derivative for each pixel in x axis and y axis;
3. Calculate the corner response function R based on formula 9;
4. Calculate partial maximum pixel  $I_{max}$  ;
5. Calculate the adaptive threshold T from formula 12.

When  $I_{max}$  is larger than the threshold T, the pixel can be treated as a corner.

### 2.3 Image Registration

After extracting feature points with the improved Harris Corner Detection algorithm, relationships between the points can be figured out by the registration algorithm. In this chapter, image registration based on corners points is in usage. The corner registration aims to find out the corresponding corner point pairs with one unique point in  $I_1$  and the other unique point in  $I_2$ . In this paper, image registration for two adjacent images is accomplished by corner registration algorithm based on the SVD(singular value decomposition) in three steps [14].

First, the cross-correlation function of regions in the two images is computed. Supposing  $I_1$  and  $I_2$  are the two adjacent images, after processing of corner detection algorithm, the two images will have M and N feature points respectively. For each feature point  $m(u_1, v_1)$  in image  $I_1$ , a  $(2p + 1) \times (2q + 1)$  rectangle region A is selected as the window function in which the feature point is the center. Rectangle region B is selected the same way for each feature point  $n(u_2, v_2)$  in image  $I_2$ . The cross-correlation function for A and B from the  $(2p + 1) \times (2q + 1)$  neighborhood constituted by the corner m in image  $I_1$  and the corner n in image  $I_2$  is defined in formula 13.

$$c(m, n) = \sum_{i=-p}^p \sum_{j=-q}^q \frac{(I_1(u_1+i, v_1+j) - \bar{A}) \times (I_2(u_2+i, v_2+j) - \bar{B})}{(2p+1) \times (2q+1) \sigma(A) \sigma(B)} \quad (13)$$

The  $\bar{A}$  means the average value of region A centering  $m$  in image  $I_1$ , and the value can be calculated in formula 14.  $\sigma(A)$  is the standard deviation of the  $(2p+1) \times (2q+1)$  neighborhood for the region A, with the value gained by formula 15.

$$\bar{A} = \overline{I_1(u_1, v_1)} = \sum_{i=-p}^p \sum_{j=-q}^q \frac{I_1(u_1+i, v_1+j)}{[(2p+1) \times (2q+1)]} \quad (14)$$

$$\sigma(A) = \sqrt{\frac{\sum_{i=-p}^p \sum_{j=-q}^q I_1(u_1+i, v_1+j)}{(2p+1) \times (2q+1)} - \overline{I_1(u_1, v_1)}} \quad (15)$$

And the  $\bar{B}$  means the average of region B centering  $n$  in  $I_2$ ,  $\sigma(B)$  is the standard deviation of the  $(2p+1) \times (2q+1)$  neighborhood for the region B.  $\bar{B}$  and  $\sigma(B)$  can be calculated in similar ways of  $\bar{A}$  and  $\sigma(A)$ .

Formula 13 illustrates the two regions' relevancy will change from -1(means that the two regions are different) to 1(means that the two regions are the same). Then the similar matrix can be calculated:

$$G(m, n) = \frac{c(m, n)+1}{2} e^{-r(m, n)^2 / 2\sigma^2} \quad (16)$$

In which  $G(m, n)$  is the Gaussian weighted distance range from 0 to 1. And  $r(m, n)$  is the Euclidean distance from  $m$  to  $n$ . Finally, the SVD is applied to  $G(m, n)$ .

$$G(m, n) = TDU^T \quad (17)$$

The  $T$  is the orthogonal matrix of  $M$  rows. The  $U$  is the orthogonal matrix of  $N$  columns. The  $D$  means the diagonal matrix of  $M$  rows and  $N$  columns.

The elements on the diagonal line of diagonal matrix  $D$  should be in descending order. The identity matrix  $E$  is constructed by setting elements which are not equal to zero to one on the diagonal line of  $D$ , so that it comes to the matrix  $P$  in formula 18.

$$P(m, n) = TEU^T \quad (18)$$

The same form of the matrix  $P$  and  $G$  makes the suited feature points stand out easily. If  $P(m, n)$  is the maximum in row as well as the maximum in column, then we consider the feature points  $m$  and  $n$  as a pair of matching points.

## 2.4 Image Fusion

As one of the core techniques in image mosaic, image fusion aims to eliminate the traces in image splicing place. In this part the weighted mean method will be used to image fusion [15]. Firstly weighted mean calculation based on gray level is applied to the feature points, and then final pixel value is determined by superimposing the pixels. Supposing the images to be mosaicked are  $f_1$  and  $f_2$ ,  $f$  is the result image, then weighted mean method can be achieved as formula 19.

$$f(x, y) = \begin{cases} f_1(x, y) & (x, y) \in f_1 \\ \omega_1 f_1(x, y) + \omega_2 f_2(x, y) & (x, y) \in (f_1 \cap f_2) \\ f_2(x, y) & (x, y) \in f_2 \end{cases} \quad (19)$$

The  $\omega_1$  and  $\omega_2$  are the weights of corresponding pixels in the overlay region of the two images. And they are conditioned by  $0 < \omega_1, \omega_2 < 1$  as well as  $\omega_1 + \omega_2 = 1$ . Though this method is simple and intuitive, fast for image fusion, can easily deal with the discrepancy of exposure, shutter appears when there exist objects in the overlay region. According to the Weber's Law, responses of the HVS (Human Visual System) for stimulus signals are based on the luminance comparison between signal and background (the average of signal) rather than on the absolute luminance. The sensitivity of HVS for gray level is in direct proportion to the logarithm of practical luminance. Usually the HVS is more sensitive to the medium gray level, and the sensitivity is more stable in this region, while nonlinear declines in two directions of high and low gray level region. So the sensitivity of human eyes is not linear as the traditional weight method considers. When the discrepancy of luminance in two images is large, it is not ideal to use the factor  $d$  in formula 20 for image fusion.

$$d = (x_2 - x)/(x_2 - x_1) \quad (20)$$

In order to solve the problem mentioned above, weight factor  $k$  is displaced with trigonometric function and the segmented weighted function.

$$d = \begin{cases} 1 - 0.5 \sin\left(\frac{x-x_1}{x_2-x_1}\right)\pi & x_1 \leq x \leq x_1 + \frac{x_2-x_1}{4} \\ \frac{(\sqrt{2}-2)(x-x_1)}{x_2-x_1} + \frac{3-\sqrt{2}}{2} & x_1 + \frac{x_2-x_1}{4} \leq x \leq x_1 + \frac{3(x_2-x_1)}{4} \\ 0.5 \sin\left(\frac{x-x_1}{x_2-x_1}\right)\pi & x_1 + \frac{3(x_2-x_1)}{4} \leq x \leq x_2 \end{cases} \quad (21)$$

### 3 Experimental Results and Discussion

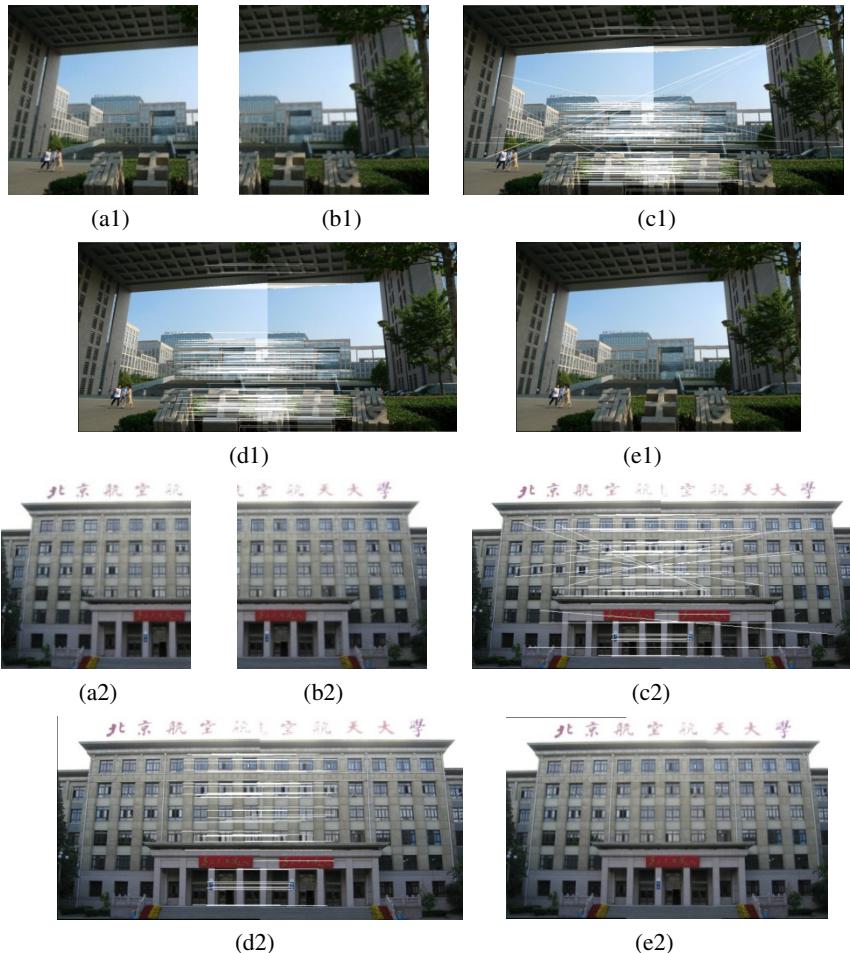
#### 3.1 Experimental Results

In this chapter the traditional method and the improved method are compared on multiple sets of images. The comparisons focus on many aspects, including effect of image mosaic, time of algorithms and the number of suited points in image registration.

Firstly the source and result images of image mosaic are showed in figure 2. For convenience of the observation, the corresponding feature points in two images are connected by a line.

In the first set of images, the (a1, b1) is the original image. Image (c1) is the result using traditional method of image registration while image (d1) is the one using the improved method. In the second set of images, the (a2, b2) is the original image. Image (c2) is the result of traditional method and (d2) uses the improved method. (e1) and (e2) are final results of image mosaic.

Secondly, five different groups of images are tested to obtain time of algorithms, including traditional image registration method and the improved one. Time of extraction and registration is measured separately. The result is shown in table 1.



**Fig. 2.** Two sets of image mosaic experiment result.

Because of the different algorithms, the number of suited point pairs and missing match point pairs are different, thus raising discrepancy on the accuracy of registration. So lastly comparisons of the number of suited points and erroneous suited points and the accuracy of registration between the two algorithms are conducted. Results are shown in table 2 and table 3.

### 3.2 Discussion

Compare to the different results, we can find out that the improved algorithm in this paper can detect the feature points more accurately and obtaining the suited features more stably. The incorrect registration in our algorithm is less than the traditional one.

Generally, the comparisons between the traditional Harris algorithm and the improved one based on the extraction time of feature points, time for registration,

number of erroneous suited points and accuracy can give us a clear look on the discrepancy of the two algorithms. And we can conclude that the improved Harris corner detection algorithm can get better result and improve the accuracy. The algorithm mentioned in our paper is effective and utility.

**Table 1.** Comparison of time for Extraction and registration

No.	Traditional algorithm			Improved algorithm		
	Extraction (Image1)/s	Extraction (Image2)/s	Registration /s	Extraction (Image1)/s	Extraction (Image2)/s	Registration /s
1	1.95	1.84	1.99	1.21	1.07	1.27
2	1.42	1.65	1.16	1.18	1.29	0.89
3	1.03	1.07	1.07	0.56	0.64	0.78
4	3.36	3.90	4.44	2.84	3.15	3.92
5	1.88	1.76	1.95	1.10	0.96	1.23

**Table 2.** Comparison on number of suited points

No.	Traditional algorithm			Improved algorithm		
	Points (Image1)	Points (Image2)	suited points	points (Image1)	points (Image2)	suited points
1	971	959	323	950	948	305
2	330	275	125	301	246	105
3	141	233	91	125	211	65
4	3015	4680	2262	2709	4299	2236
5	709	830	201	654	689	112

**Table 3.** Comparison on the erroneous suited points and accuracy

No.	Traditional algorithm		Improved algorithm	
	erroneous suited points	Accuracy	erroneous suited points	Accuracy
1	62	81.0%	46	85%
2	21	83.2%	14	86.7%
3	14	85.1%	6	91.1%
4	189	91.6%	156	93.0%
5	18	91.0%	6	94.4%

## 4 Conclusion

For serious clustering phenomenon and fewer corner points in the texture region of traditional Harris corner detection algorithm, this paper proposed an improved adaptive threshold setting algorithm by calculating the corner response function of second-order, avoiding the impact of the scale factor k and the threshold T on corner detection. For obvious traces of weighted average method in the jointing places, this paper proposed a new weighted average method based on trigonometric functions for

image fusion. Experimental results show that the modified algorithm can effectively eliminate the gap generated by image mosaic, and improve stitching accuracy and speed.

## References

1. Peleg, S., Herman, J.: Panoramic mosaics by manifold projection. In: Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 338–343. IEEE (June 1997)
2. Liu, J., Li, L.: A Brief Introduction of Reconstruction Technology of Distant Medical Consultation for Large Image(远程医疗会诊中拼接大型医学图像技术简介). China Contemporary Medicine 6(10), 62 – 63 (2000)
3. Nie, S.D., Si, J.Y.: Methodological Study of Automatically Mosaicing for Medical Microscopic Images(医学显微图像自动拼接的方法研究). Chinese Journal of Biomedical Engineering 24(2), 173 - 178 (2005)
4. Cai, Y., Hu, X.: Short wave infrared imaging technology and its defense application (短波红外成像技术及其军事应用). Infrared and Laser Engineering 35(6), 643 – 647 (2006)
5. Wen, H. Y.: Creating image-based VR using a self-calibration fisheye lens (遥感图像拼接算法研究). Doctoral Dissertation, 华中科技大学 (2009)
6. Xiong, Y., Turkowski, K.: Creating image-based VR using a self-calibrating fisheye lens. In: Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 237–243. IEEE (June 1997)
7. Wang, J., Shi, J., Wu, X.X.: Survey of image mosaics techniques. Application Research of Computers 25(7), 1940–1943 (2008)
8. Szeliski, R.: Image alignment and stitching: A tutorial. Foundations and Trends® in Computer Graphics and Vision 2(1), 1–104 (2006)
9. Li, Q., Zhang, B.: A fast matching algorithm based on image gray value. Journal of Software 17(2), 216–222 (2006)
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, p. 50 (1988)
11. Gumustekin, S., Hall, R.W.: Mosaic image generation on a flattened Gaussian sphere. In: Proceedings 3rd IEEE Workshop on Applications of Computer Vision, WACV 1996, pp. 50–55. IEEE (December 1996)
12. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 31(4), 532–540 (1983)
13. Noble, J.A.: Finding corners. Image and Vision Computing 6(2), 121–128 (1988)
14. Feng, Y.P., Dai, M.: An Image Mosaic Algorithm Based on Corner Features (一种基于角点特征的图像拼接融合算法). Microelectronics & Computer (7), 21 - 23 (2009)
15. Szeliski, R.: Video mosaics for virtual environments. IEEE Computer Graphics and Applications 16(2), 22–30 (1996)

# A Study of Ancient Ceramics Verification Based on Vision Methods

Yunqi Tang<sup>1,\*</sup>, Jianwei Ding<sup>2</sup>, and Wei Guo<sup>1</sup>

<sup>1</sup> Department of Criminal Science and Technology, People's Public Security University of China, Beijing, 100038, China

<sup>2</sup> Department of Police Information Engineering, People's Public Security University of China, Beijing, 100038, China  
`{tangyunqi,dingjianwei,guowei1}@ppsuc.edu.cn`

**Abstract.** Ceramics appraisal is a hot topic in field of cultural relic collection. Traditionally, there are mainly two types of ceramics appraisal methods, which are experience-based methods and technology-based methods. In practice, the both methods would cause high cost and time consuming. In this paper, a novel vision based method, which is mainly inspired by the idea of biometrics recognition techniques, is proposed to achieve efficiently verification of the identity of a ceramics. In this method, the microscopic information of a ceramics captured by a digital microscope camera are used as the characteristics for verification. In technical detail, SURF(Speeded Up Robust Features) is first employed to align the probe image to the gallery images. LBP(Local Binary Patterns) features are then extracted from the two aligned images. Finally, Chi-square distance is calculated to measure the similarity between probe and gallery. Experiments on the dataset constructed by this paper demonstrate the state-of-the-art performance of our method.

**Keywords:** Ancient ceramics verification, SURF, Local binary patterns, Biometric recognition.

## 1 Introduction

Ceramic, which owns significant archaeological value, is an important source for all archaeological studies and for studies on art and technology of material culture. Ancient ceramic is usually valuable, and great many fake ceramics are illegal produced for economic purpose. How to verify the identity of a ceramic is a critical problem for the circulation of ceramics.

Traditionally, appraisal certificate is used as the identity card of a ceramic. It can efficiently indicate the identity of a ceramic. However, the appraisal certificate of a ceramic is usually produced by an authentication institution or authenticator and is easy to be counterfeited. Thus, in most transactions of ancient ceramics, the appraisal certificate is not convincing, and a live appraisal process is needed. Currently, there are two types of ceramic appraisal methods, which are

---

\* Corresponding author.

experience-based methods and technology-based methods. In experience-based methods, appraisal results are subjectively generated according to the experience of experts. While in technology-based methods, physical or chemical technology means are utilized to achieve the appraisal purpose. Generally, both types of appraisal method would cause high cost and time consuming.

In this paper, a novel ceramic verification method is presented to achieve efficient appraisal of ceramic based on the idea of biometric verification. In this method, the characteristic of a ceramic, such as bubble distribution, decoration et.al, is utilized to form its digital identity, which can be stored in its appraisal certification. Due to the reason that the digital identity of a ceramic only relays on the ceramic itself, it is difficult to counterfeit the digital identity of a ceramic. In practice, once a person claims to initiate a ceramic transaction, the digital identity of the ceramic is live generated to compare with the identity stored in the appraisal certification. If the live generated identity matches the identity stored in appraisal certification, then the ceramic can be regarded as a genuine. Otherwise, the ceramic would be regarded as a fake. With this method, it is easy and efficient to verify the identity of a ceramic. This would significantly decrease the cost of ceramics transaction.

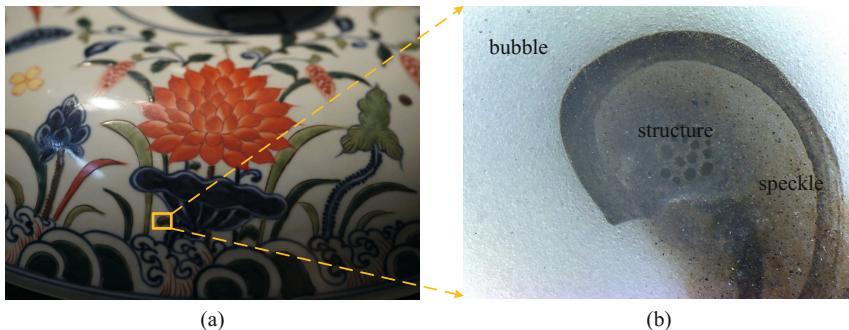
The main contribution of this paper lies two fold. Firstly, vision based techniques are introduced to achieve live ceramic verification. To the best of our knowledge, it is the first time that vision based method is applied to address the problem of ceramic verification. Secondly, an efficient ceramic verification method is presented in this paper. SURF (Speeded Up Robust Features) is employed to align two ceramic images, and LBP (Local Binary Patterns) features are extracted from ceramic images for generating the digital identity of a ceramic.

The rest of this paper is organized as follows. In section 2, we analyze the possibility of using vision based method to verify the identity of a ceramic and describe the key problems of this method. Section 3 provides technical details of the proposed vision-based ancient ceramic verification algorithm. Section 4 shows the experimental results. And finally, this paper is concluded in section 5.

## 2 Problem Analysis

In biometric recognition, any human physiological and/or behavioral characteristic can be used as a biometric characteristic as long as it satisfies the requirements of universality, distinctiveness, permanence and collectability [1]. Actually, the characteristics of a ceramic satisfies the above four requirements.

Fig.1 shows the microscopic characteristics of a ceramic. The image shown in Fig.1 (a) is the surface of a ceramic captured by a digital camera. And the image shown in Fig.1 (b), which is captured by a digital microscope camera, is the microscopic information of (a). We can see that the microscopic characteristics contained in a ceramic's surface, such as bubble, speckle and geometric pattern et.al, are distinctive. The reason is that the microscopic information of a ceramic is directly decided by the materials prescription, processing technology,



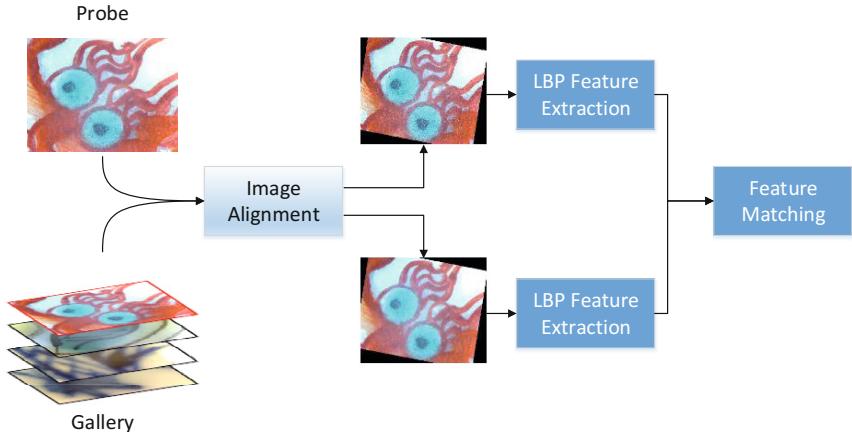
**Fig. 1.** Microscopic characteristics of a ceramic. a) A ceramic image. b) The microscopic characteristics of the ceramic shown in (a). It is captured by a digital microscope camera.

and sinter environment. The materials prescription and processing technology are easy to reproduce, while the sinter environment is difficult to duplicate. This makes the detail information contained in the surface of a ceramic distinctiveness. Furthermore, it is obvious that the microscopic characteristics of ceramics are universal, permanent and collectable. Thus, biometrics recognition techniques, which follow the process of preprocessing, feature extraction and feature matching, can be used for ceramic verification.

- Preprocessing. The preprocessing step is usually used to normalize the input data before features are extracted. For example, in face recognition, a face image are normalized to certain illumination and pose condition. However, in the case of ceramic verification, the preprocessing problem is quite different. Firstly, there is not a unified object (such as face, iris, hand et.al.) contained in ceramic images. No uniform normalizing target is existed. Secondly, the microscopic characteristics are captured by a digital microscope camera. Its illumination condition is controllable. Thus, the main problem lies in preprocessing step is how to align the images captured from the same ceramic.
- Feature extraction and matching. Theoretically, any image representation model can be used for representing ceramic images. Whereas, different representation model would achieve different performance. To decide the representation model of ceramic images is the other critical problem of ceramic verification.

### 3 Vision Based Ceramic Verification Method

In this section, we introduce the ceramic verification method proposed by this paper. The framework of the proposed method is designed as Fig.2. In this method, the registered gallery images are directly stored as image format instead



**Fig. 2.** The framework of proposed method

of feature format. When a test image is probing to the gallery images, alignment techniques are first employed to normalize the probe image with the gallery images. LBP(Local Binary Patterns) are then extracted from the probe image and gallery images. Finally, the distance between the two sets of LBP features are calculated for measuring the similarity between probe image and gallery images.

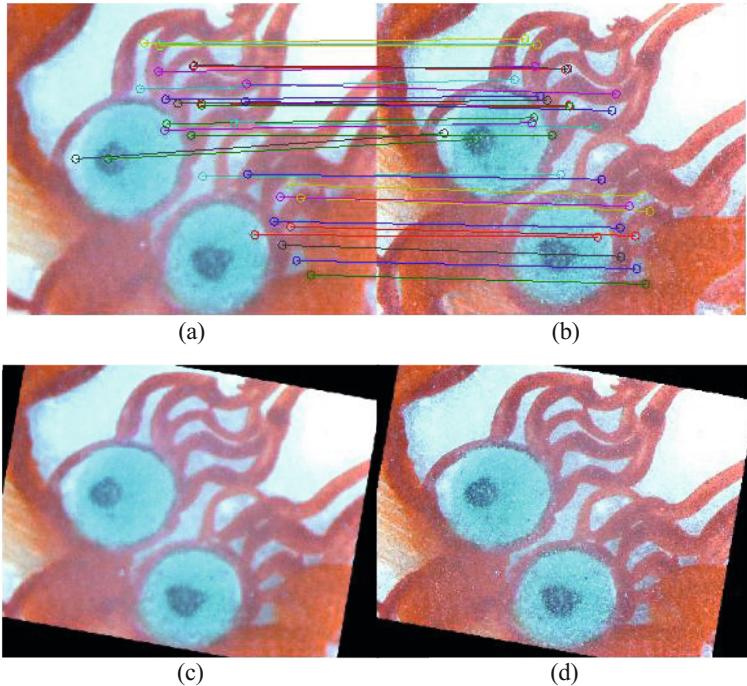
### 3.1 Image Alignment

Image alignment technique is a good choice for ceramic images normalization due to the reason that no uniform normalizing target is contained in ceramic images. In this paper, SURF(Speeded Up Robust Features) is used to perform the task of image alignment.

SURF is a scale and in-plane rotation invariant feature proposed by Herbert Bay [2,3]. It has been widely used in object recognition and 3D reconstruction. The idea of SURF is partly inspired by the SIFT(Scale Invariant Feature Transform) descriptor [4,5], which follows the process of detection and description. Detectors are first employed to find the interest points in an image, and then the descriptors are used to extract the feature vectors at each interest point. Differently, Hessian-matrix approximation operating on the integral image are used in SURF to locate the interest points instead of difference of Gaussians (DoG) filter used in SIFT. And the first-order Haar wavelet responses in x and y directions are used as the descriptor in SURF instead of the gradient is used by SIFT. Due to the two improvements, SURF is faster and more robust than SIFT.

The preliminary experimental results of image alignment using SURF algorithm are shown in Fig.3. The two images shown in Fig.3(a) are the images captured from the same point of the same ceramic with different viewpoints.

The image shown in Fig.3(b) is the left image of Fig.3(a) transformed to the viewpoint of right image. From this figure, we can see that SURF can achieve accurately alignment of two ceramic images.



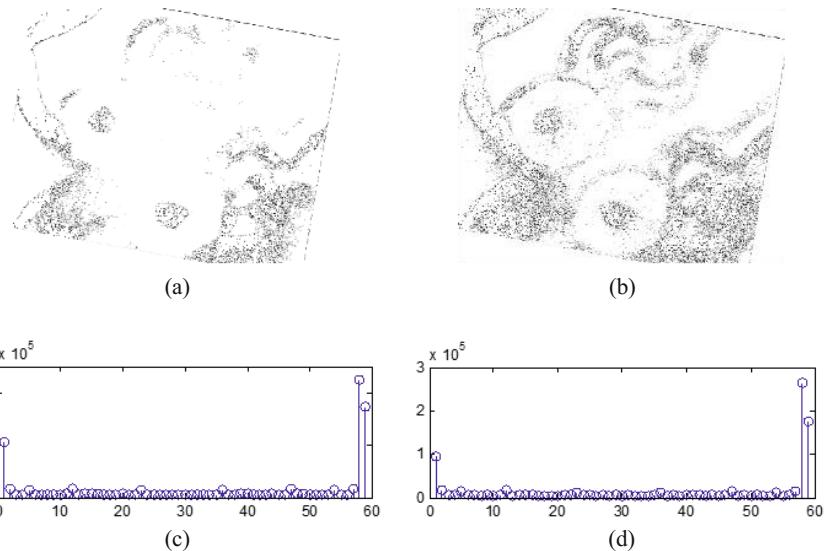
**Fig. 3.** Alignment of two ceramic images using SURF algorithm. a) Key points matching results of SURF algorithm. b) Projection of the left image to the view of the right image.

### 3.2 Feature Extraction and Matching

In this subsection, we introduce how is a ceramic image represented and how is the similarity of two ceramic images measured.

As shown in the Fig.1, the microscopic characteristics contained in a ceramic image mainly include bubble, speckle and geometric pattern. All of these microscopic information results in the distinctive texture of a ceramic image. Thus, texture descriptor is suitable for ceramic images representation. In this paper, we employ LBP (Local Binary Patterns) to represent the texture information contained in the ceramic image.

Local Binary Pattern is a simple yet very efficient texture operator. The original LBP operator is introduced by Ojala et al [6]. Its basic idea is to label the pixels of an image by thresholding the neighborhood of each pixel with the value of the center pixel and consider the result as a binary number. Then the histogram of



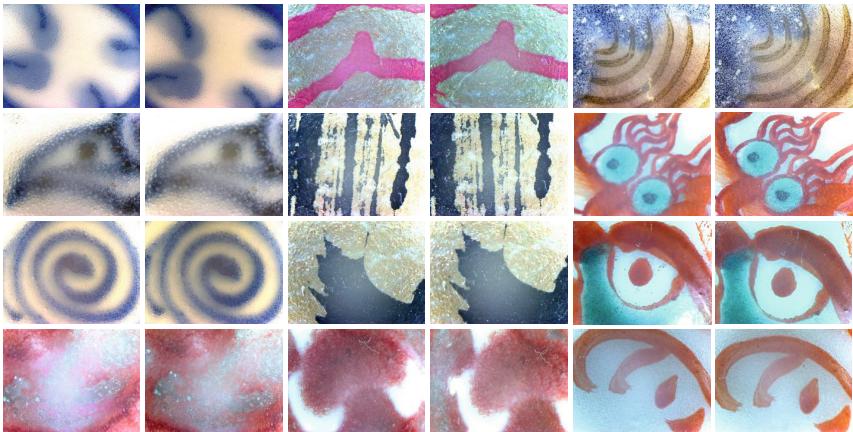
**Fig. 4.** LBP images of ceramic images. a) shows the LBP image of Fig.3 c). b) shows the LBP image of Fig.3 d). c) is the histogram of image a). d) is the histogram of image b).

the labels is used as a texture descriptor. Later the operator was extended to use neighbourhoods of different sizes [7] by using circular neighbourhoods instead of 3x3-neighbourhood. With this improvement, the extended local binary patterns allow any radius and number of pixels in the neighbourhood. LBP operators can be regarded as a unifying approach to the traditionally divergent statistical and structural models of texture analysis. One advantage of LBP operator is its robustness to monotonic gray-scale changes, such as illumination variations. The other important property of LBP operator is its computational simplicity, which makes it possible to analyze images in challenging real-time settings. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications [8].

The LBP images and histograms of two ceramic images are shown in Fig.4. From this figure, we can see that the LBP histograms of two ceramic images, which are captured from the same point of the same ceramic, are similar with each other. The Chi-square measure of the two LBP histograms are 1.02.

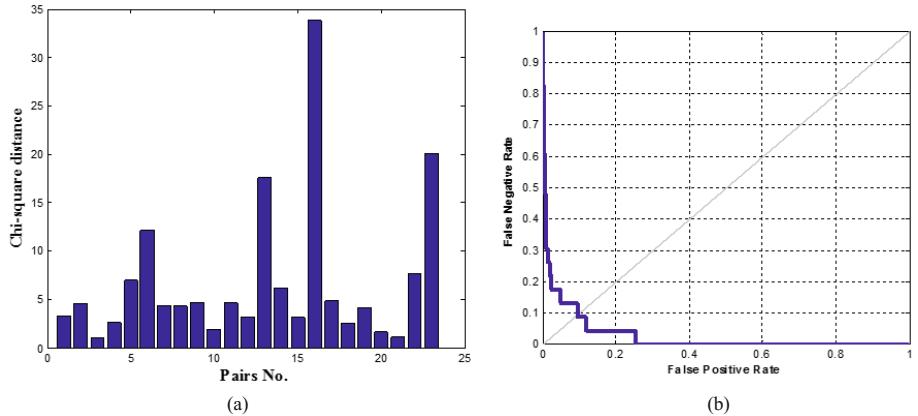
## 4 Experiments and Results

To evaluate the performance of proposed method, we constructed a ceramic database using a digital microscope. There are totally 23 pairs of ceramic images within this database. The two paired images are captured from the same ceramic with different viewpoints. Each ceramic image is stored as JPEG format with



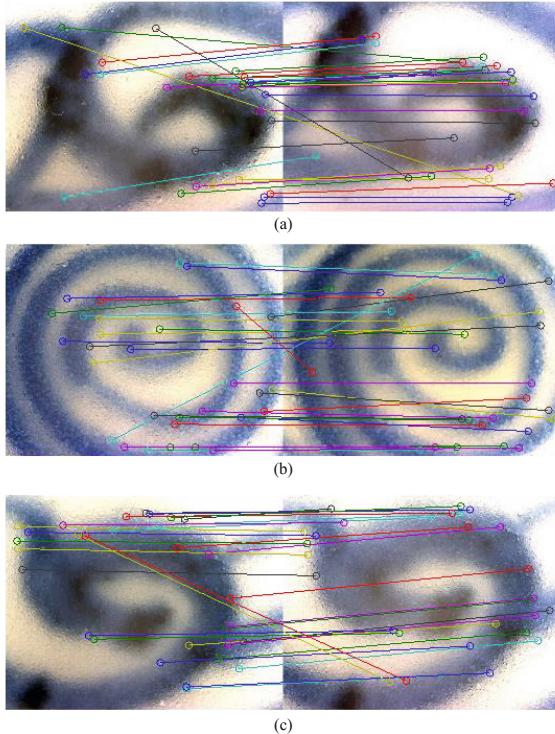
**Fig. 5.** Paired ceramic image samples of the constructed database

a resolution of 3200\*2400. Fig.5 shows paired ceramic samples of this database. To the best of our knowledge, it is the first ceramic image dataset.



**Fig. 6.** The experimental results on the constructed dataset. a) shows the Chi-square distance of intraclass. b) shows the ROC curve of the experimental results.

Fig.6 shows the experimental results on the constructed database. The Chi-square distance of intra-class is presented in Fig.6(a). We can see that the Chi-square distance of most intra-class samples are less than 10. On the contrast, the average Chi-square distance of interclass is 58.8. And the ROC curve of the experimental results is shown in Fig.6(b). The equal error rate (EER) of this method is 0.087. This demonstrates that the proposed method can achieve good performance for ceramic verification.



**Fig. 7.** The alignment results of failure samples. a) the alignment result of the 13rd paired ceramic images whose Chi-square distance is 17.6. b) the alignment result of the 16th paired ceramic images whose Chi-square distance is 33.8. c) the alignment result of the 23rd paired ceramic images whose Chi-square distance is 20.1.

As shown in the Fig.6(a), the Chi-square distance of the 13rd, 16th, and 23rd pairs of ceramic images is respectively 17.6, 33.8 and 20.1, which are significantly larger than other paired samples. The reason for this phenomena mainly lies in the alignment algorithm employed by this method. Fig.7 shows outputs of the SURF algorithm used in the alignment process. Obviously, there are errors within the matched key points generated by SURF algorithm, which will cause further errors to image alignment. This demonstrates that more accurate alignment will improve the accuracy of this method.

## 5 Conclusions

This paper presents a promising method for ceramic verification. The main novelty of this paper is design of a vision based method for ceramic verification. Firstly, SURF (Speeded Up Robust Features) is employed to align the probe image to the gallery images. Secondly, LBP (Local Binary Patterns) features

are extracted from the two aligned images. Finally, Chi-square distance is calculated to measure the similarity between probe and gallery. Experiments on constructed databases have demonstrated good performance of our method.

**Acknowledgments.** The author thanks the ancient ceramics samples of Shiquan Liu. This work is supported by the Fundamental Research Funds for the Central Universities, and the Joint Development Project for the Central Universities in Beijing.

## References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14(1), 4–20 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features(SURF). *Computer Vision and Image Understanding* 110(3), 346–359 (2008)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
4. Lowe, D.G.: Object recognition from local scale-invariant features. In: *International Conference on Computer Vision(ICCV)*, pp. 1150–1157 (1999)
5. Lowe, D.G.: Distince image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
6. Ojala, T., Pietikainen, M., Harwood, D.A.: comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* 29(1), 51–59 (1996)
7. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
8. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J(G.) (eds.) *ECCV 2004*. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)

# A Two-Stage Blind Image Color Correction Using Color Cast Estimation

Dawei Zhu<sup>1,2</sup>, Li Chen<sup>1,2</sup>, Jing Tian<sup>1,2</sup>, and Xiaotong Huang<sup>1,2</sup>

<sup>1</sup>School of Computer Science and Technology, Wuhan University of Science and Technology,  
Wuhan, China, 430081

<sup>2</sup>Hubei Province Key Laboratory of Intelligent Information Processing  
and Real-time Industrial System, Wuhan University of Science and Technology,  
Wuhan, China, 430081  
{971919309}@qq.com

**Abstract.** The color cast images usually have serious loss of the color information and are inconvenient for visual observation and image analysis. To tackle this problem, a novel two-stage image color cast correction scheme is proposed in this paper. Firstly, the proposed approach performs the color cast and stable channel detection by using extreme intensity ratio of the original image. Second, the distorted image color is restored by solving the constrained problem with the degree of color variation and the above-detected color cast and stable channel. The experimental results using surveillance videos demonstrate that the proposed scheme is not only feasible but also effectively. In addition, the results satisfy human subjective perception as well.

**Keywords:** Color cast image, stable channel detection, color cast estimation, blind color correction.

## 1 Introduction

Color cast means the variation between the collected images and the real color of object surface. It is usually incurred by the influence of the environmental light source, reflection characteristics of the object, and the parameters of acquisition device. This color variation may introduce undesirable effects for human perception (e.g., police investigation). Moreover, it may negatively affect the performance of computer vision methods for different applications such as object recognition, tracking [1-2].

The aim of color correction is to eliminate the effect of the color cast. Over the past decade, a variety of algorithms are used to color correction for color cast images and videos. The *Shades of Grey* (SoG) approach [3] introduces Minkowski-norm into Grey World algorithms. It utilizes the Minkowski-norm distance to replace the simple averaging method. Although they are conceptually simpler, it can not realize color correction for the original image that does not matching the grey hypothesis. The Grey Edge hypothesis in [4] from observation of the distribution of image color derivative in opponent color space is a relative regular

ellipse, and the axis of the ellipse and the direction of light are same. Then, this method assumes the average of the reflectance differences in a scene is achromatic to realize color correction. However, the Grey Edge method gets insufficient color restoration for badly color cast images, due to the hypothesis is invalid. Color correction in multi-scaled retinex [5] is proposed using a modified local average image to improve the color rendition and reduce the color distortion by the dominant chromaticity of the original image. Due to the use of complex image processing techniques, it is time-consuming. The survey of many recent developments and state-of-the-art color constancy methods is presented in [6]. Its main focus is on the estimation of the illuminant using a single image from a regular digital camera. Due to use regular digital camera, it does not result in satisfactory results for surveillance videos. Spectral reflectance estimation method in [7] is proposed to utilize images including both near-infrared image and visible light image to realize spectral reflectance estimation, and then achieves color correction. However, near infrared imaging system usually have higher prices than traditional visible cameras. This disadvantage limits the scope of the application.

This paper studies blind color correction, where the input image is not known to be color cast. For that, a novel color correction scheme is proposed in this paper. The core of the proposed scheme lies in seeking a stable channel and achieving color correction by the guided of this stable channel. The proposed approach consists of two stages. First, the input color cast image is preprocessed for obtaining extreme intensity ratio. And then the proposed approach performs the color cast detection and stable channel detection. Second, difference intensity ratio is calculated to estimate the degree of color variation. And then, the bind color correction is realized.

The rest of this paper is organized as follows. First, the stable color channel detection is proposed in Section 2. The proposed color correction approach is proposed in Section 3. Experimental results are presented in Section 4. Finally, Section 5 concludes this paper.

## 2 Stable Channel Detection

Various methods have been proposed to achieve color correction. It is important to note that the reference image cannot be obtained from video surveillance systems. It leads to a blind problem, which is difference from the conventional assumptions for most of color correction methods. The proposed scheme is focus on blind surveillance video images.

Three types of *color variation* (CV) are presented to describe the bias of color channels in an original image. They are (i) additive variation, (ii) subtractive variation, and (iii) stable. The first type represents that the intensity of one color channel is larger than other color channels. The second type represents that the intensity of one color channel is smaller than other color channels. The third type represents that the intensity among color channels is balance.

A red cast image is shown in Fig. 1(a). The Fig. 1(b) shows the histogram of the original image. The curve with dotted line represents red channel, the curve with solid

line represents green channel, and the curve with point and line represents blue channel. It can be found that the intensity of blue is smaller than other two colors, and the intensity of red is larger than other two colors. Based on the description of color variation, the proportion of bright pixels and dark pixels in each channel are calculated as follows.

Considering an image with RGB color space, the bright channel information  $\mathbf{H}_{bri}$  and the dark channel information  $\mathbf{H}_{dark}$  of image  $\mathbf{I}$  are defined as:

$$\mathbf{H}_{bri} = \max(\mathbf{I}^R, \mathbf{I}^G, \mathbf{I}^B), \mathbf{H}_{dark} = \min(\mathbf{I}^R, \mathbf{I}^G, \mathbf{I}^B) \quad (1)$$

where  $\mathbf{I}^R, \mathbf{I}^G, \mathbf{I}^B$  represent red, green and blue component of image  $\mathbf{I}$ , respectively.  $\max(\cdot)$  and  $\min(\cdot)$  denotes the operator of maximum and minimum. The *extreme intensity ratio* (EIR) can be calculated as:

$$EIR_i^c = S_i^c / (H \times W), c \in \{R, G, B\} \quad (2)$$

where  $H$  and  $W$  represent the height and width of image  $\mathbf{I}^c$ .  $EIR_i^c$  represents the bright and dark intensity ratio of RGB channels. Moreover,  $S_i^c \in \{S_{bri}^c, S_{dark}^c\}$  is defined as:

$$S_i^c = \sum_{\mathbf{x}} \mathbf{L}_i^c(\mathbf{x}) \quad (3)$$

$$\mathbf{L}_i^c(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{I}^c(\mathbf{x}) = \mathbf{H}_i(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

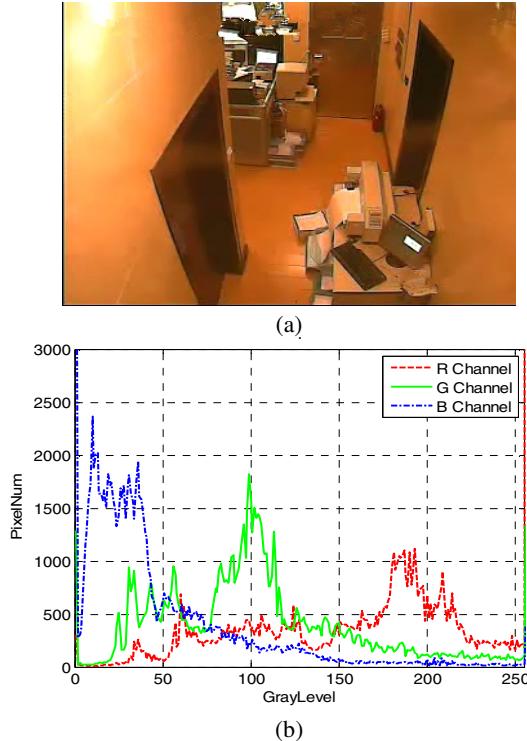
where  $\mathbf{x}$  represents spatial position in the image,  $\mathbf{I}^c(\mathbf{x})$  represent the gray value in pixel  $\mathbf{x}$ ,  $\mathbf{L}_i^c(\mathbf{x})$  is the symbol of bright or dark pixel in  $\mathbf{x}$ ,  $S_i^c$  represents the total number of pixels that satisfy the conditions.

Defining  $V_{bri}^{st}$ ,  $V_{bri}^{nd}$  and  $V_{dark}^{st}$ ,  $V_{dark}^{nd}$  respectively as the 1<sup>st</sup> and 2<sup>nd</sup> maximum value of  $EIR_{bri}^c$  and  $EIR_{dark}^c$ . The CV of each color channel can be determined as follows:

$$\mathbf{I}^c \in \begin{cases} \text{additive,} & \text{if } V_{bri}^{st} - V_{bri}^{nd} \geq T_1, EIR_{dark}^c = \min(EIR_{dark}) \\ \text{subtractive,} & \text{if } V_{dark}^{st} - V_{dark}^{nd} \geq T_2, EIR_{bri}^c = \min(EIR_{bri}) \\ \text{stable,} & \text{otherwise} \end{cases} \quad (5)$$

where  $T_1$  and  $T_2$  are pre-defined threshold. If the three channels are all stable, the original image does not have color cast. The additive channel is that its intensity is larger than other two channels. The subtractive channel is that its intensity is smaller than other two channels. Otherwise, the channel is stable.

In Fig. 1(b), the intensity distribution of blue channel and red channel arise the polarization phenomenon. On the basis of equation (5), the blue is subtractive type, and the red channel is additive type, and the green channel is stable type.



**Fig. 1.** Example of color variation. (a) Original image. (b) Histogram of the image (a).

### 3 Color Correction

According to section 2, the stable channel is detected. The color correction for the color variation channels based on the stable channel is proposed in this section. It is based on a simple conception, using the color channel with more reliable information as a guider to restore the other channels [8].

Inspired by the image degradation model proposed in [9], which is formulated as follows:

$$g(\mathbf{x}) = f(\mathbf{x})t(\mathbf{x}) + A(1-t(\mathbf{x})) \quad (6)$$

where  $g$  represents the original image,  $f$  represents the scene radiance,  $t(\mathbf{x})$  is the medium transmission, and  $A$  is the global atmospheric light. The proposed scheme uses the degradation model (6) to color correction.

Let  $\psi(\mathbf{x}) = 1 - 1/t(\mathbf{x})$ ,  $q(\mathbf{x}) = A - g(\mathbf{x})$ , it gives

$$f(\mathbf{x}) = g(\mathbf{x}) + q(\mathbf{x})\psi(\mathbf{x}) \quad (7)$$

To achieve this end, the unknown term  $q\psi$  in (7) need to be estimated. In the following, how to estimate the parameters as well as obtain the restored image is presented.

Considering the variation of color channels, it can have large or small bias. The key is to find some features to classify the serious or slight situation. It can be perceived that there are differences among the intensity of color channels through observation defined as *difference intensity* (DI). It can be calculated as follows:

$$DI^c(\mathbf{x}) = \sum_{\mathbf{x}} (\Delta^c(\mathbf{x})) / S_{bri}^c \quad (8)$$

where  $S_{bri}^c$  represents the number of bright pixels using equation (3) and (4), respectively.  $\Delta^c(\mathbf{x})$  can be computed as:

$$\Delta^R(\mathbf{x}) = \Delta\mathbf{RG}(\mathbf{x}) + \Delta\mathbf{RB}(\mathbf{x}) \quad (9)$$

where  $\Delta^R(\mathbf{x})$  denotes the intensity difference between red channel and the rest channels,  $\Delta\mathbf{RG}(\mathbf{x})$  and  $\Delta\mathbf{RB}(\mathbf{x})$  defined as:

$$\Delta\mathbf{RG}(\mathbf{x}) = \begin{cases} \mathbf{I}^R(\mathbf{x}) - \mathbf{I}^G(\mathbf{x}), & \text{if } \mathbf{I}^R(\mathbf{x}) = \mathbf{H}_{bri}(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

$$\Delta\mathbf{RB}(\mathbf{x}) = \begin{cases} \mathbf{I}^R(\mathbf{x}) - \mathbf{I}^B(\mathbf{x}), & \text{if } \mathbf{I}^R(\mathbf{x}) = \mathbf{H}_{bri}(\mathbf{x}) \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

It can be extended to G and B color channels. After that, the *difference intensity ratio* (DIR) can be obtained as follows:

$$DIR^c = \frac{1}{1 + e^{-(DIR^c - a)/b}} \quad (12)$$

where  $a$  and  $b$  represent pre-parameter and can self-adapting select based on the luminance of image.

For the aim of color correction, the *degree of color variation* (DCV) is presented to quantize the intensity variation, defined as  $\alpha_{DCV}$ . It is calculated as follows.

$$\alpha_{DCV} = \sum_{c \in \{R, G, B\}} (EIR_{bri}^c \times DIR^c) \quad (13)$$

where  $\alpha \in [0, 1]$ ,  $EIR_{bri}$  and  $DIR$  denotes the bright intensity ratio and DIR, respectively.

First, considering the property of DCV, assuming the combination of the DCV and the stable channel as the estimation of  $q\psi$ . That is

$$q\psi \approx \alpha_{DCV} p \quad (14)$$

where  $p$  represents the guided matrix. Then,  $f$  can be solved from  $g$  in equation (7) by solving the optimization problem stated in equation (15) using conjugate gradient method with initialized  $f_0 = g$  as follows:

$$\hat{f} = \arg \min_f \sum_{i \in c} \|f_i - g_i - q_i \psi_i\|^2 + \lambda \|w(\text{sgn} \cdot \Delta f_i)\|^2 \quad (15)$$

where  $\Delta f = f_i^n - f_i^{n-1}$  denotes difference between the current iterative result and the previous iterative result.  $\text{sgn}$  is a sign function.

$$w(K) = \begin{cases} 0, & K \leq 0 \\ \beta, & K > 0 \end{cases} \quad (16)$$

$$\text{sgn} = \begin{cases} 1, & \text{if } f_i \text{ is additive} \\ -1, & \text{if } f_i \text{ is subtractive} \\ 0, & \text{if } f_i \text{ is stable} \end{cases} \quad (17)$$

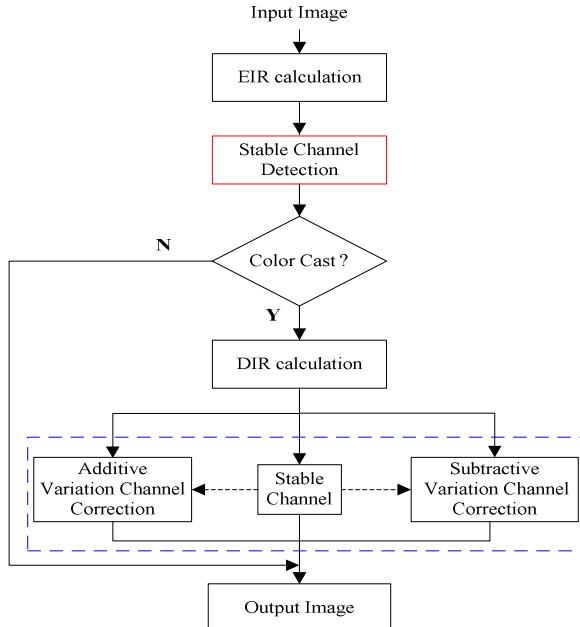
where  $\beta$  denotes penalty factor, and  $f_i$  denotes the current result.

The overall framework of the proposed algorithm is summarized in Fig. 2 and described as follows.

**Step1:** Given an input image, calculate EIR by equation (2), and determine color variation of each channel by equation (5). Finally, achieve stable channel and color cast detection basing on color variation.

**Step2:** Calculate DIR by equation (12), and then obtain the DCV by equation (13);

**Step3:** Perform color variation correction by solving the constrained equation (15) based on the estimation of unknown parameters  $q\psi$  in equation (14).

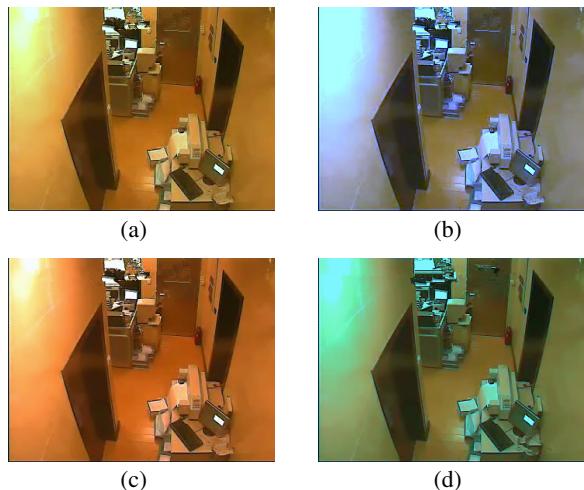


**Fig. 2.** The overview of the proposed approach

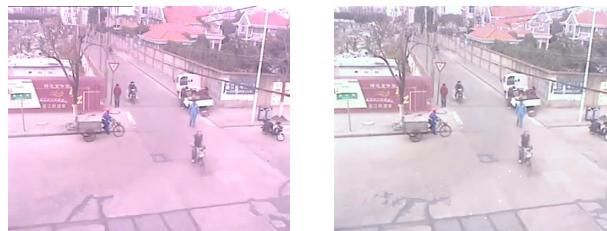
## 4 Experiments

The proposed method is assessed by testing it on several surveillance videos. In the experiment, the logo and time are masked for privacy. The  $T_1$ ,  $T_2$  used in the proposed approach are set to be 0.6,0.6. It is worthy mentioning that the proposed method is designed for blind surveillance video images for real-life applications.

Fig. 3 compares various experimental results of image (in Fig. 1(a)) obtained by the proposed method, the Grey Edge method [4], the Grey World method [1], the Max-RGB method [2] are respectively presented. As seen in Fig. 3, the proposed approach achieves more natural image result for human perception. For example, the wall in Fig. 3(a) and Fig. 3(c) displays yellow and orange, and in Fig. 3 (b), the wall is white cast. By contrast, the result of proposed scheme is more comfortable for human visual perception. In addition, the fire extinguisher in the corner is also better.



**Fig. 3.** Various corrected color image (Fig. 1(a)). (a)-(c). Results of [4, 1, 2]. Respectively; (d) Proposed method.



**Fig. 4.** Color correction of the proposed method. The first column is the input image, and the second column is the output image.



**Fig. 4. (Continued)**

Fig. 4 shows the results of other surveillance videos by the proposed scheme. It can be observed the proposed method significantly restores the color of the images. It almost does not have excessive corrections, and also has a well visual effect with human subjective evaluation.

## 5 Conclusions

In this paper, a novel two-stage scheme for color correction is proposed. The proposed scheme uses the extreme intensity ratio to process the color cast and stable channel detection, and then restores color by solving the constrained problem based on the degree of color variation and above prior color cast and stable channel detection. Experimental results show that the proposed method is capable of achieving color correction and improves the perceptual quality.

**Acknowledgment.** This work was supported by National Natural Science Foundation of China (No. 61105010, 61375017), Program for Outstanding Young Science and Technology Innovation Teams in Higher Education Institutions of Hubei Province, China (No. T201202).

## References

- [1] Buchsbaum, G.: A spatial processor model for object colour perception. *Journal of the Franklin Institute* 310(1), 1–26 (1980)
- [2] Finlayson, G.D., Hordley, S.D., Hubel, P.M.: Color by correlation: A simple, unifying framework for color constancy. *IEEE Trans. Pattern Analysis and Machine Intelligence* 23(11), 1209–1221 (2001)

- [3] Finlayson, G.D., Trezzi, E.: Shades of gray and colour constancy. In: The 12th Color and Imaging Conference, Scottsdale, USA, vol. 12, pp. 37–41 (2004)
- [4] Weijer, J.V., Gevers, T., Gijsenij, A.: Edge-based color constancy. *IEEE Trans. Image Processing* 16(9), 2207–2214 (2007)
- [5] Jang, I.S., Park, K.H., Ha, Y.H.: Color correction by estimation of dominant chromaticity in multi-scaled retinex. *Journal of Imaging Science and Technology* 53(5), 1–11 (2009)
- [6] Gijsenij, A., Gevers, T., Weijer, J.: Computational color constancy: survey and experiments. *IEEE Trans. Image Processing* 20(9), 2475–2489 (2011)
- [7] Igarashi, Y., Ogawa, T., Haseyama, M.: Spectral reflectance estimation from visible light components and near-infrared components. In: Proc. IEEE Int. Conf. Image Processing, Australia, vol. 2013, pp. 2388–2392 (2013)
- [8] He, K., Sun, J., Tang, X.: Guided image filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part I*. LNCS, vol. 6311, pp. 1–14. Springer, Heidelberg (2010)
- [9] Koschmieder, H.: Theorie der horizontale Sichtweite. *Journal of Beitr. Phys. Freien Atmos.* 12, 171–181 (1924)

# Encoding Optimization Using Nearest Neighbor Descriptor

Muhammad Rauf, Yongzhen Huang, and Liang Wang

National Lab of Pattern Recognition,  
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

**Abstract.** The Bag-of-words framework is probably one of the best models used in image classification. In this model, coding plays a very important role in the classification process. There are many coding methods that have been proposed to encode images in different ways. The relationship between different codewords is studied, but the relationship among descriptors is not fully discovered. In this work, we aim to draw a relationship between descriptors, and propose a new method that can be used with other coding methods to improve the performance. The basic idea behind this is encoding the descriptor not only with its nearest codewords but also with the codewords of its nearest neighboring descriptors. Experiments on several benchmark datasets show that even using this simple relationship between the descriptors helps to improve coding methods.

**Keywords:** Nearest neighbor descriptor, Group saliency coding, Soft coding, Local constraint linear coding.

## 1 Introduction

One of the most important research areas in computer vision is image classification. There are different kinds of techniques used to serve this purpose. All these techniques have their benefits and drawbacks. Some work well on one kind of dataset and others can perform better on other kind of datasets. In all these techniques, the most commonly used framework is the Bag-of-words framework (BoW)[1][2]. This model consists of several steps, which starts from feature extraction and ends with classification. The hierarchy of these steps is, after feature extraction a codebook is generated and followed by feature coding, and before the classification feature pooling is performed.

All these steps have their own importance in the whole process of image classification using BoW. In recent years, encoding attracts lots of attention. There are different kinds of encoding methods that have been introduced to get better performance. Recent work[4][5] show that different coding methods perform different, even under the same framework. Soft voting outperforms hard voting[1] and the fisher kernel[6] has better performance than soft voting [3] with the same number of code words. These three are voting based methods and if we compare these voting based methods with reconstruction based coding[4],

like local constraint linear coding (LLC)[7], we find that LLC has better results than the voting based coding methods. On the other hand the saliency[8] and group saliency coding[9] methods have implementation advantages over the reconstruction based coding, and perform faster than LLC. There are other coding methods introduce to improve the performance e.g., Laplacian sparse coding[10], multi-layer group sparse coding[11], improved Fisher kernel coding[12], Local tangent-based coding methods[13] and many more.

One thing that is common in all these methods is to encode one descriptor with codewords. In this process, we exactly do not know the relationship between a descriptor and its adjacent descriptors. If the descriptor extraction is not very dense then what are the influence of one descriptor to its neighboring descriptors and their codewords, i.e., the codewords used to encode descriptors. The main focus of our work is, to encode the descriptor by using the nearest neighbor descriptor's (NND) codewords and observe the change in performance. We explore a relationship between descriptors and by using this relationship, we update the codewords of descriptors. Our proposed technique is very simple and easy to implement.

The rest of the paper is arranged as follows. In Section 2 we introduce our proposed method in detail. In Section 3 first we discuss the datasets and the coding methods, and afterwards we evaluate our proposed technique. At the end in Section 4 we present the conclusion and our future work.

## 2 Nearest Neighbor Descriptor

The proposed method not only considers the structure of K-nearest codewords to a descriptor, but also takes account of the structure of neighboring descriptor codewords. We present a new technique that uses the descriptor-to-descriptor relationship during the encoding process. Results show that the locality of the descriptors has a very important role in encoding.

Our implementation is done in two different phases. First, we find K-nearest codewords of a descriptor and finally we update each descriptor's codewords based on the NND codewords. Let  $X = [x_1, x_2, \dots, x_N] \in R^D \times N$  be  $N$   $D$ -dimensional descriptor form an image, and  $B = [b_1, b_2, \dots, b_M] \in R^D \times M$  be a codebook with  $M$  codewords.

### 2.1 Local Code Assignment

In this phase, we encode the descriptor with  $K$  codewords using the existing encoding methods.  $K$  is set to be a small number[20] and  $[b_1, b_2, \dots, b_K]$  is  $K$  closest codewords of  $x$  e.g.,  $K=3$  in Fig. 1(a). This is the local assignment of the nearest codewords to the descriptor. In the next phase, we generate new codewords that is based on the descriptor's and its neighboring descriptor's codewords.

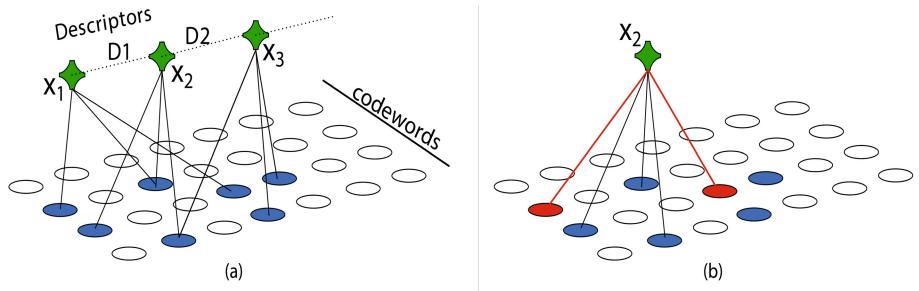
## 2.2 Nearest Neighboring Descriptor

The position of the descriptor from its neighboring descriptor has an important factor during encoding. We update codewords for every descriptor by using its codewords and codewords of its NND. Suppose  $Y_i$  is the set of codewords of descriptor  $x_i$  and  $Z_i$  is the set of codewords of NND of  $x_i$ . We choose the codewords which are used to encode  $x_i$  by using Equation 1:

$$Y'_i = Y_i \cup (Z_i \setminus Y_i), \quad (Z_i \setminus Y_i) = \{b \in Z_i | b \notin Y_i\} \quad (1)$$

where  $Y'_i$  is the updated set of codewords of the descriptor  $x_i$  and ' $\setminus$ ' stands for the relative complement function.  $(Z_i \setminus Y_i)$  represents the relative complement of  $Y_i$  in  $Z_i$ , the set of codewords that are presented in  $Z_i$  but not in  $Y_i$ .

Our method is illustrated in Fig. 1. First, we find the nearest neighboring descriptor and then we assign the new codewords to the descriptor according to the nearest neighboring descriptor's codewords. Suppose we are going to encode  $x_2$ . The first step is to find the NND of  $x_2$ . Consider  $D_1$  and  $D_2$  are two distances between  $x_2$  to  $x_1$  and  $x_2$  to  $x_3$  respectively. Suppose  $D_1$  is less than  $D_2$ , so  $x_1$  is the nearest neighboring descriptor of  $x_2$ . By using the equation 1, we assign new codewords to  $x_2$ .

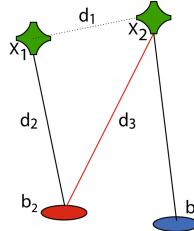


**Fig. 1.** Code selection on the base of the nearest neighboring descriptor

The distance between codewords and descriptors plays a very important role in the encoding process. The next step is to find the distance of codewords to it's new descriptor. After assigning new distance, we select  $K$  nearest codewords.

Suppose  $b_1$  and  $b_2$  are the codewords of the descriptor  $x_2$  as shown in Fig. 2, where  $b_2$  is new codeword of  $x_2$  from it's NND. Suppose  $d_1$  is the descriptor to descriptor distance and  $d_2$  is the distance of codeword to it's original descriptor. We need to calculate  $d_3$ , the distance of the codeword to its new descriptor.

We use a simple technique to estimate the new distance of codeword to it's new descriptor. We use Pythagorean theorem[14] to calculate the distance. This will not get the exact distance but it will proximate the distance and improve the speed. According to our observation this estimation error is negligible with



**Fig. 2.** Distance measurement of codewords to descriptors

respect to the fast performance of our technique. By using equation 2 we calculate  $d_3$ .

$$d_3 \approx \sqrt{d_1^2 + d_2^2} \quad (2)$$

In the final stage we supply these codewords to the selected encoding method to finalize the encoding process.

### 3 Experiments and Discussion

#### 3.1 Datasets

The following four datasets are used for experimental study.

**Scene 15**[15]. There are 4,485 images in the scene 15 dataset and these images belong to 15 different categories, each of which contains 200 to 400 images. We randomly select 100 images for training and the remaining for testing.

**Caltech 101**[16]. This dataset contains 9,145 images from 101 different categories. These categories contain from 31 to 800 different numbers of images. We use the standard setting for this dataset.

**VOC2007**[17]. There are 9,963 images in this dataset distributed into 20 classes. These images vary in their size, scales, viewpoint and other image properties. These images are divided into training and testing sets. VOC2007 is one of the major datasets used in image classification.

**UIUC Sports**[18]. The UIUC Sport dataset consists of 1,574 sports images belonging to 8 different categories. We use this dataset for extensive study of our proposed technique.

#### 3.2 Experimental Setting

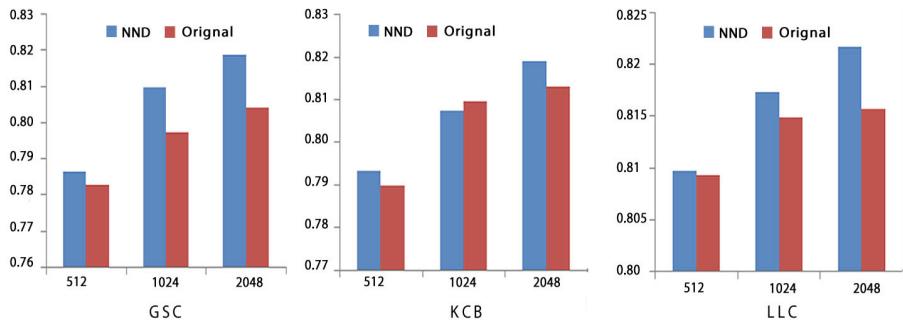
We use three different encoding methods from three different encoding classes to observe the performance of our technique, i.e., kernel codebook encoding (KCB)[19], locality constrained linear coding (LLC) and group saliency coding (GSC). For all these methods the codeword size  $K$  is 5 and the codebook sizes are set to 512, 1024 and 2048. We use SIFT descriptor[21] for all these experiments.

The evaluation is performed with two different experimental settings. First, we evaluate performance with three different datasets and feature extraction of image is with 10 step size (i.e., extracting a descriptor over every 10 pixels). We use Scene 15, Caltech 101 and VOC2007 datasets for these experiments. In second group of experiments we use the UIUC-Sports dataset to evaluate the performance with 8, 10, 15 and 20 step size of image feature extraction.

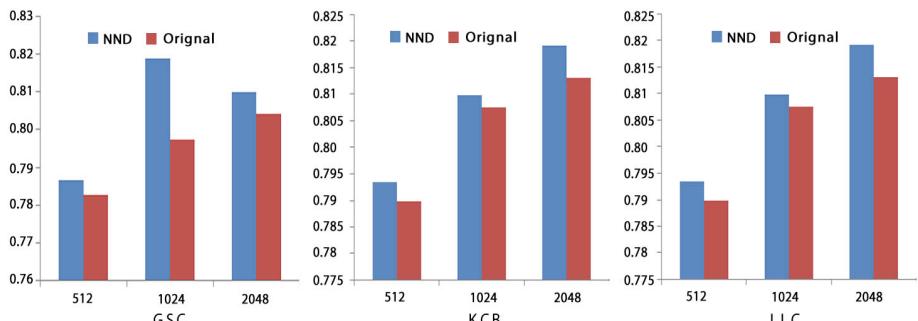
### 3.3 Basic Results

As mentioned above in these experiments, we use three different datasets with one feature extraction size. Results of Caltech 101, Scene 15 and VOC2007 are shown in Fig. 3, Fig. 4 and Fig. 5 respectively. Each figure contains the results by GSC, KCB and LLC. The results of our proposed technique and original methods are compared. These results suggest that the performance of the NND based method is increased.

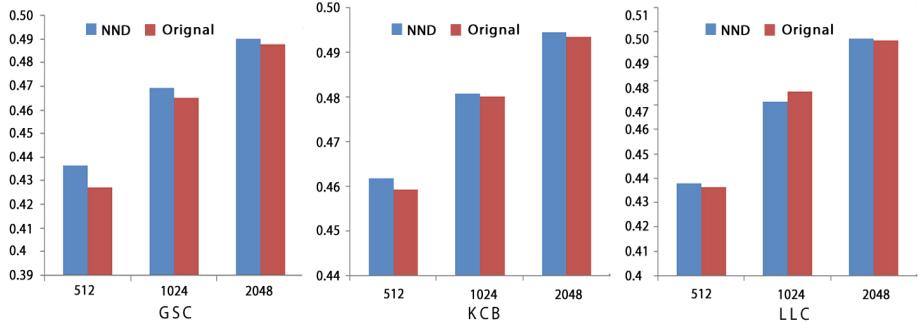
These different charts from each group involve three different codebook sizes. From these accuracy bars, it is clear that the accuracy is improved after using



**Fig. 3.** Experimental results on the Caltech 101 Dataset



**Fig. 4.** Experimental results on the Scene 15 Dataset



**Fig. 5.** Experimental results on the VOC2007 Dataset

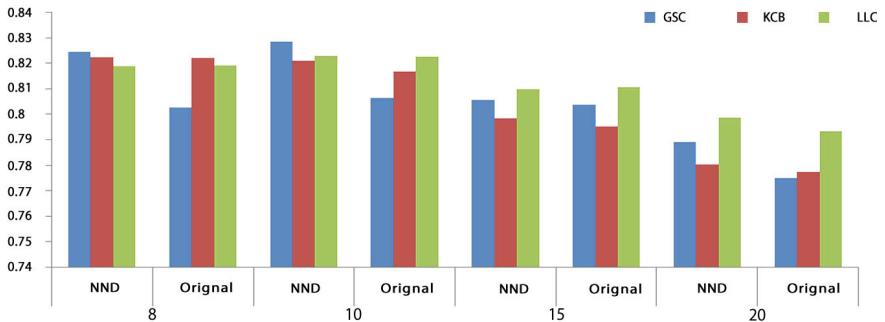
our technique. Although the improvement is not very large in some cases but still it has a small change in the performance.

We can observe that even the property of relationship is simple, it is still able to perform well. It should be noted that if we are able to explore a good relationship between the descriptors, we may obtain more improvement. These results show that NND performs with persistent enhancement on different datasets with different encoding methods.

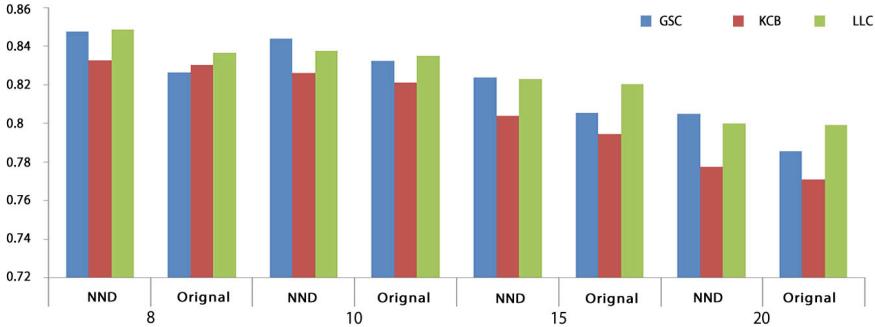
### 3.4 Different Sampling Rate Evaluation

For further testing our proposed technique, we use UIUC-Sports dataset with different feature extraction sizes. In these experiments we use 20, 15, 10 and 8 step size of image feature extractions respectively.

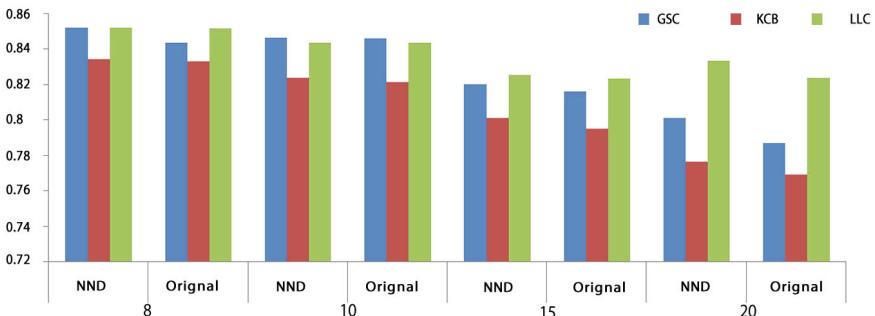
We evaluate the performance of our proposed technique by comparing with original version of GSC, KCB and LLC. The results shown in Fig. 6, Fig. 7 and Fig. 8 give us clear observation on the performance improvement. Our technique again performs better in all these experimental settings. The performance dif-



**Fig. 6.** Experimental results on the UIUC-Sport Dataset with a codebook size 512



**Fig. 7.** Experimental results on the UIUC-Sport Dataset with a codebook size 1024



**Fig. 8.** Experimental results on the UIUC-Sport Dataset with a codebook size 2048

ference of the NND and the original method is large when size of descriptors are not very dense. In low density of descriptor, the distance between the descriptor is large, so encoding with our proposed technique has more clear effects. This is probably because with the low descriptor density, descriptors are more scattered than with high descriptor density.

## 4 Conclusion and Future Work

In this paper, we have developed a new technique to improve the existing methods via exploring the relationship between descriptors. Our work has shown that if the relationship between the descriptors are developed in a meaningful way, it can help to get better results in terms of image classification. We have used this technique with GSC, KCB and LLC, and obtained improvement in all evaluation conditions.

Our future work is to extend this technique to video classification. It is believed that this method will generate better performance in video classification due to the finding that our proposed technique has better performance with low descriptor density, which is usually the case in video classification based on the bag-of-words framework.

**Acknowledgments.** This research is founded by National Basic Research Program of China (Grant No. 2012CB316300) and National Natural Science Foundation of China (Grant No.61175003).

## References

1. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV, (2004)
2. van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
3. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(7), 1271–1283 (2010)
4. Yu, K., Zhang, T., Gong, Y.: Nonlinear Learning using Local Coordinate Coding. In: NIPS (2009)
5. Huang, Y., Wu, Z., Wang, L., Tan, T.: Feature coding in image classification: A comprehensive study. IEEE Trans. Pattern Anal. Mach. Intell. 36(3), 493–506 (2014)
6. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
7. Huang, Y., Huang, K., Yu, Y., Tan, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
8. Huang, Y., Huang, K., Yu, Y., Tan, T.: Salient coding for image classification. In: CVPR (2011)
9. Wu, Z., Huang, Y., Wang, L., Tan, T.: Group encoding of local features in image classification. In: ICPR (2012)
10. Gao, S., Tsang, I., Chia, L., Zhao, P.: Local features are not lonely - laplacian sparse coding for image classification. In: ECCV (2010)
11. Gao, S., Chia, L.T., Tsang, I.W.: Multi-layer group sparse coding for concurrent image classification and annotation. In: CVPR (2011)
12. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
13. Yu, K., Zhang, T.: Improved local coordinate coding using local tangents. In: ICML (2010)
14. [http://en.wikipedia.org/wiki/Pythagorean\\_theorem](http://en.wikipedia.org/wiki/Pythagorean_theorem)
15. (2006), [http://www-cvr.ai.uiuc.edu/ponce\\_grp/data/scene\\_categories/scene\\_categories.zip](http://www-cvr.ai.uiuc.edu/ponce_grp/data/scene_categories/scene_categories.zip)
16. [http://www.vision.caltech.edu/Image\\_Datasets/Caltech101/101\\_ObjectCategories.tar.gz](http://www.vision.caltech.edu/Image_Datasets/Caltech101/101_ObjectCategories.tar.gz)
17. <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/index.html>
18. [http://vision.stanford.edu/lijiali/event\\_dataset/event\\_dataset.rar](http://vision.stanford.edu/lijiali/event_dataset/event_dataset.rar)
19. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR (2008)
20. Liu, L., Wang, L., Liu, X.: In defense of softassignment coding. In: ICCV (2011)
21. David, G.L.: Distinctive image features from dcaleinvariant key-points. International Journal of Computer Vision 2(60), 91–110 (2004)

# Multi-modal Image Fusion with KNN Matting

Xia Zhang, Hui Lin, Xudong Kang, and Shutao Li

Vision and Image Processing Labrataory, College of Electrical and Information Engineering, Hunan University, Changsha, China, 410082  
{xiazhang2013,xudong\_kang,shutao\_li}@hnu.edu.cn,  
linhui1965@126.com

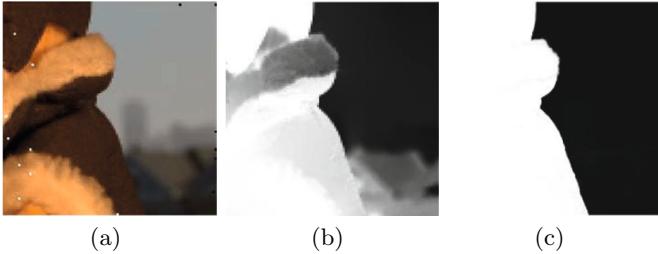
**Abstract.** A single captured image of a scene is usually insufficient to reveal all the details due to the imaging limitations of single sensor. To solve this problem, multiple images capturing the same scene with different sensors can be combined into a single fused image which preserves the complementary information of all input images. In this paper, a novel K nearest neighbor (KNN) matting based image fusion technique is proposed which consists of the following steps: First, the salient pixels of each input image is detected using a Laplician filtering based method. Then, guided by the salient pixels and the spatial correlation among adjacent pixels, the KNN matting method is used to calculate a globally optimal weight map for each input image. Finally, the fused image is obtained by calculating the weighed average of the input images. Experiments demonstrate that the proposed algorithm can generate high-quality fused images in terms of good visual quality and high objective indexes. Comparisons with a number of recently proposed fusion techniques show that the proposed method generates better results in most cases.

**Keywords:** Image fusion, KNN matting, Laplician filtering, weighted average.

## 1 Introduction

Image fusion is able to fuse the complementary information preserved in different images of the same scene, so as to obtain a single fused image which provides more comprehensive information. The fused images are usually more useful for human and machine perception, and thus, image fusion has been widely applied for digital photography, object detection and related applications [1].

Recently, a large number of image fusion methods have been proposed such as multiscale image fusion and optimization based fusion. Multiscale fusion aims at proposing different multi-scale coefficients and novel fusion rules to guide the fusion of coefficients [2,3]. Since the multi-scale coefficients provide an accurate representation of images, these methods can well preserve the details of different images [2]. However, since spatial information is not considered, multi-scale based methods cannot ensure the color and brightness consistency of the fused image [4]. To solve this problem, optimization based image fusion approaches, e.g., generalized random walks [5], and matting based method [6] have been proposed



**Fig. 1.** An example of image matting. (a) Input images and clicks. (b) Alpha matte obtained by the closed form matting method [7]. (c) Alpha matte obtained by the KNN matting method [8].

for fusion of multi-exposure and multi-focus images. These methods first estimate accurate weights by solving an energy function. The source images are then fused together by weighted average of pixel values. In [6], robust matting has been successfully applied for fusion of multi-focus images. However, this method relies heavily on the accurate estimate of initial weights. More importantly, the method is designed only for multi-focus image fusion.

In order to extend image matting for fusion multi-modal images captured by different imaging sensors, a novel KNN matting based fusion method is proposed in this paper. Experimental results show that the proposed method gives a performance comparable with state-of-the-art fusion approaches including the traditional robust matting based fusion method [6].

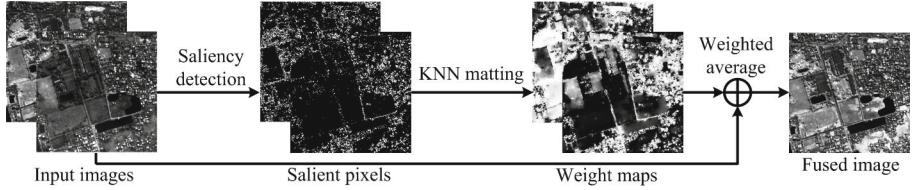
## 2 Image Matting

Image matting is a technique to accurately distinguish foreground objects from background [7]. Specifically, an input image  $I$  can be represented as a combination of foreground color  $F$  and background color  $B$ .

$$I = \alpha \times F + (1 - \alpha) \times B \quad (1)$$

where  $\times$  represents the multiplication operation element by element,  $\alpha$  ranging from 0 to 1 is the foregrounds opacities which is usually named as the alpha matte. The objective of image matting is to calculate the alpha matte  $\alpha$ , the foreground color  $F$ , and the background color  $B$  from a single image  $I$ . Obviously, this problem is under constrained which has infinite solutions. Therefore, in most cases, user inputs and prior assumptions are required in addition to the original image.

Fig. 1 gives an example of image matting. Taking Fig. 1(a) as the inputs which contain the input image, and user clicks (black point indicates that this pixel is a background pixel, white point indicates that this pixel is a foreground pixel), different matting algorithms are able to obtain different matting results shown in Fig. 1(b) and (c). It can be seen that, although there has only few user



**Fig. 2.** A schematic of the proposed KNN matting based image fusion method

clicks as inputs, the KNN matting method [8] is still able to obtain an accurate alpha matte which can distinguish the foreground object from the background. By contrast, the closed form matting method [7] fails in detecting the accurate foreground object in this example. Therefore, the KNN matting is expected to be a more suitable solution for the proposed matting based image fusion method.

### 3 Proposed Method

Fig. 2 summarizes the main processes of the proposed KNN matting based fusion method (GFF). First, an laplacian filter is utilized to detect the salient pixels in each input image. Then, taking the detected salient pixels and the input images as the inputs of KNN matting, the weight maps  $\alpha$  used for image fusion can be estimated. Finally, the fused image is obtained by weighted average of input images.

#### 3.1 Salient Pixels Detection

In order to detect the salient pixels of each input image, Laplacian filtering is applied to each source image  $I_n$  to obtain the filtered image  $H_n$ .

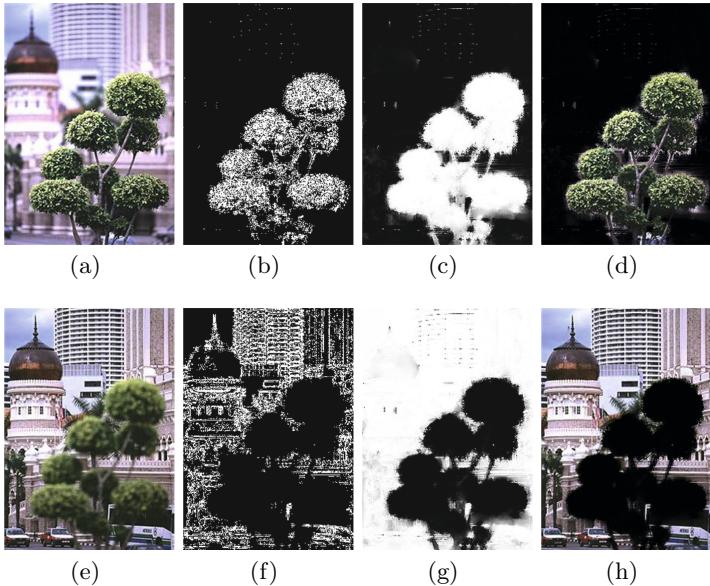
$$H_n = I_n * L \quad (2)$$

where  $L$  is a  $3 \times 3$  Laplacian filter,  $I_n$  refers to the  $n$ th input image. Then, these pixels give much higher salience values are assigned as salient pixels.

$$S_n^i = \begin{cases} 1 & \text{if } H_n^i - \max_{m,m \neq n} H_m^i > \sigma, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

where  $i$  refers to the  $i$ th pixel,  $\sigma$  is a free threshold to control the number of detected salient pixels. In this paper, the parameter  $\sigma$  is fixed as 0.1 for all experiments.

Fig. 3(c) and (d) shows the detected salient pixels of Fig. 3(a) and (b), respectively. It can be seen that the major function of the Laplician filter is to find the rough positions of the salient objects in each input image.



**Fig. 3.** (a) and (e) Input images. (b) and (f) Salient pixels detected by Laplician filtering. Weight maps  $\alpha_1$  (c) and  $\alpha_2$  (g) estimated by the KNN matting method [8]. Accurate salient objects of the input images obtained by  $\alpha_1 \times I_1$  (d) and  $\alpha_2 \times I_2$  (h).

### 3.2 KNN Matting and Fusion

After salient pixel detection, through considering the spatial consistency between adjacent pixels, KNN matting is used to estimate the accurate weight map for image fusion. Specifically, the detected salient pixels are considered as the input clicks of the KNN matting algorithm. The detailed description of the KNN matting method can be found in [8]. Here, the KNN matting step is represented using a KNN function as follows.

$$\alpha_n = \text{KNN} \{I_n, S_n\} \quad (4)$$

For this function, the inputs are the input images  $I_n$  (see Fig. 3(a) and (e)), and the salient pixel maps  $S_n$  (see Fig. 3(b) and (f)). The outputs are the resulting alpha matte  $\alpha_n$  for each input image  $I_n$  (see Fig. 3(c) and (g)). As shown in Fig. 3, the salient object in each input image can be detected accurately by calculating the multiplication between the alpha matte and the corresponding input image pixel by pixel. Therefore, the fused image can be easily constructed by calculating a weighed sum of the input images with the alpha mattes serving as weight maps as follows.

$$F = \sum_{n=1:N} \alpha_n \times I_n \quad (5)$$

where  $N$  is the number of input images.

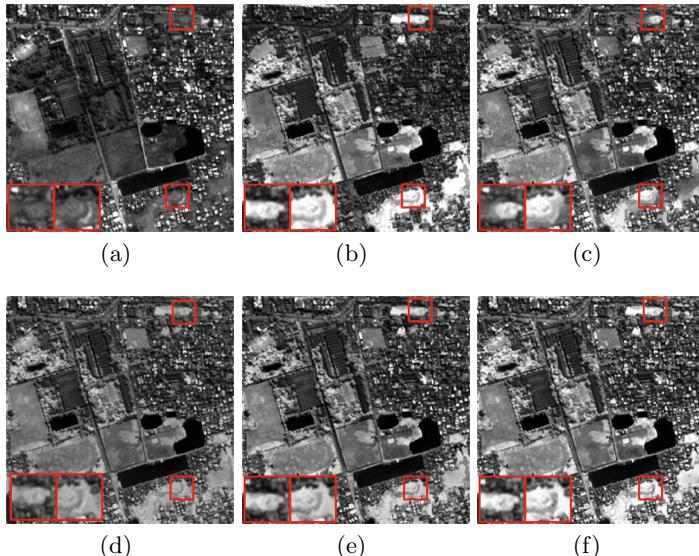
## 4 Experiments

### 4.1 Experimental Setup

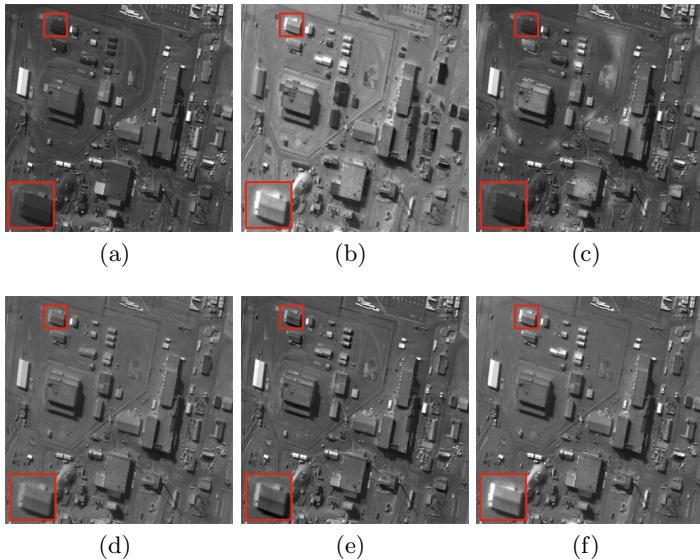
In this section, the proposed KNN based fusion method (KNNF) is compared with three recently proposed image fusion methods, i.e., the robust matting based multifocus fusion method (RMMF) [4], the guided filtering based fusion method (GFF) [6], and the recursive filtering based fusion method (RFF) [9]. These methods are implemented by using the default parameters given by the corresponding authors. For the proposed KNNF method, the default parameters given in [8] is adopted for the KNN matting algorithm. Furthermore, four widely used image fusion quality metrics, i.e.,  $Q_0$ ,  $SSIM$ ,  $Q_w$ , and  $MI$  have been used to test the objective fusion performance of different methods. A detailed description of these quality metrics can be found in [9].

### 4.2 Fusion Results

The first experiment is performed on the multispectral images shown in Fig. 4(a) and (b), where the two images are two different bands of a multispectral image. In order to compare the performance of different methods clearly, the details in the input images and the fused images are magnified for close-up comparison. As shown in Fig. 4, the RMMF method may produce false details which are not existed in the input images. Although the RFF and GFF methods are



**Fig. 4.** Input multi-spectral images and fused images obtained by different fusion methods. (a) and (b) Input images. Fused images obtained by the RMMF method (c), RFF (d), GFF(e), and proposed KNNF method (f).

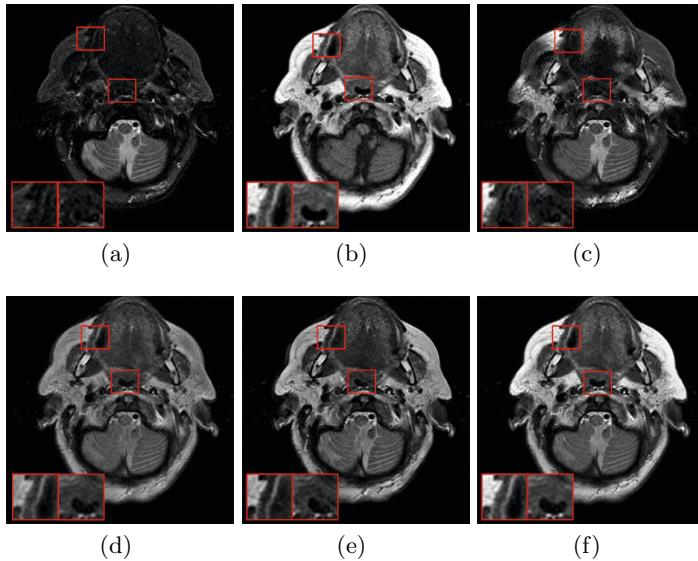


**Fig. 5.** Input multi-spectral images and fused images obtained by different fusion methods. (a) and (b) Input images. Fused images obtained by the RMMF method (c), RFF (d), GFF(e), and proposed KNNF method (f).

able to preserve the complementary information of input images, image contrast may be decreased in the fused images especially for the result obtained by the RFF method. By contrast, the proposed KNNF method can well preserve most of useful information in the input images without decrease the contrast of the fused image. Table 1 compares the objective performances of different methods. As shown in this table, the proposed KNNF method gives the best fusion performance for the first experiment in terms of the highest value of all four quality indexes.

Another multispectral image fusion example is presented in Fig. 5. This figure shows that the proposed method is able to preserve most useful information in the input images. By contrast, the results generated by RMMF may lose some important image details. The RFF and GFF methods may decrease the local contrast of fused images. Similar to the first experiment, table 1 also shows that the proposed KNNF method gives the best objective performances for the second experiment.

The third experiment is performed on multi-modal medical images. Two medical images are captured by using different sensors and thus able to reveal different types of details about a human's brain. Fig. 6 shows that the RMMF, RFF, and GFF methods may fail in detecting some salient brain structures. By contrast, the proposed KNNF can preserve the salient information of different images. Furthermore, Table 1 shows that the proposed KNNF method gives the best performance in terms of the highest value of most quality indexes, except ranking as third for the  $Q_0$  Metric.



**Fig. 6.** Input multi-modal images and fused images obtained by different fusion methods. (a) and (b) Input images. Fused images obtained by the RMMF method (c), RFF (d), GFF(e), and proposed KNNF method (f).

**Table 1.** Objective Performances of Different Image Fusion Methods Measured by Four Objective Quality Indexes

Experiment-1				
Metrics	RMMF	RFF	GFF	KNNF
$Q_0$	0.7315	0.7047	0.7679	<b>0.7707</b>
$SSIM$	0.8099	0.5668	0.7455	<b>0.8984</b>
$Q_w$	0.7274	0.698	0.773	<b>0.779</b>
$MI$	0.3239	0.2979	0.2995	<b>0.3513</b>
Experiment-2				
Metrics	RMMF	RFF	GFF	KNNF
$Q_0$	0.4975	0.5885	0.6095	<b>0.6136</b>
$SSIM$	0.6346	0.4921	0.6764	<b>0.6791</b>
$Q_w$	0.6045	0.6451	0.7046	<b>0.7248</b>
$MI$	0.4104	0.3342	0.4035	<b>0.4154</b>
Experiment-3				
Metrics	RMMF	RFF	GFF	KNNF
$Q_0$	0.4206	0.5476	<b>0.5596</b>	0.5463
$SSIM$	0.5442	0.5142	0.7006	<b>0.8316</b>
$Q_w$	0.3738	0.5733	0.6588	<b>0.7893</b>
$MI$	0.3038	0.3041	0.3054	<b>0.3482</b>

## 5 Conclusions

A novel image fusion method based on KNN matting is presented in this paper. The proposed method utilizes the Laplician filter to detect the salient pixels of each source image, which is simple and effective. More importantly, the KNN matting is used to make full use of the spatial correlations between neighborhood pixels for weight estimation. Experiments show that the proposed method can well preserve the complementary information of multiple input images captured by different sensors. The proposed method can achieve competitive subjective and objective performances compared with several recently proposed image fusion methods. Therefore, it will be quite useful in real applications.

**Acknowledgment.** This work was supported in part by the National Natural Science Foundation of China under Grant No. 61172161, the National Natural Science Foundation for Distinguished Young Scholars of China under Grant No. 61325007.

## References

1. Goshtasby, A.A., Nikolov, S.: Image fusion: Advances in the state of the art. *Inf. Fusion.* 8, 114–118 (2007)
2. Looney, D., Mandic, D.: Multiscale image fusion using complex extensions of EMD. *IEEE Trans. Signal Process.* 57, 1626–1630 (2009)
3. Pajares, G., Cruz, J.M.: A wavelet-based image fusion tutorial. *Pattern Recognit.* 37, 1855–1872 (2004)
4. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. *IEEE Trans. Image Process.* 22, 2864–2875 (2013)
5. Shen, R., Cheng, I., Shi, J., Basu, A.: Generalized random walks for fusion of multi-exposure images. *IEEE Trans. Image Process.* 20, 3634–3646 (2011)
6. Li, S., Kang, X., Hu, J., Yang, B.: Image matting for fusion of multi focus images in dynamic scenes. *Inf. Fusion.* 14, 147–162 (2013)
7. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intel.* 30, 228–242 (2008)
8. Cheng, Q., Li, D., Tang, C.: KNN matting. *IEEE Trans. Pattern Anal. Mach. Intel.* 35, 2175–2188 (2013)
9. Li, S., Kang, X.: Fast multi-exposure image fusion with median filter and recursive filter. *IEEE Trans. Consum. Electronics* 20, 3634–3646 (2011)

# A Two-Step Adaptive Descreening Method for Scanned Halftone Image

Fei Chen<sup>1</sup>, Shutao Li<sup>1</sup>, Le Xu<sup>2</sup>, Bin Sun<sup>1</sup>, and Jun Sun<sup>3</sup>

<sup>1</sup>College of Electrical and Information Engineering, Hunan University,  
Changsha, China, 410082

<sup>2</sup>School of Automation, Southeast University, Nanjing, China, 210096

<sup>3</sup>Information Department, Fujitsu Research and Development Center,  
Beijing, China, 100025

[cs.fei9009@hotmail.com](mailto:cs.fei9009@hotmail.com), [shutao\\_li@hnu.edu.cn](mailto:shutao_li@hnu.edu.cn),  
[522608533@qq.com](mailto:522608533@qq.com), [sunbinxs@126.com](mailto:sunbinxs@126.com), [sunjun@cn.fujitsu.com](mailto:sunjun@cn.fujitsu.com)

**Abstract.** Halftoning is a necessary technique for electrophotographic printers to print continuous tone images. Scanned images obtained from such printed hard copies are corrupted by screen like artifacts called halftone patterns. Descreening aims to recover high quality continuous tone image from the scanned image. In this paper, a two-step descreening method is proposed to remove screen like artifacts adaptively. Firstly, an Extreme Learning Machine (ELM) based halftone image classification scheme is introduced to categorize the scanned images into different resolutions. Then in the halftone pattern removal step, patch similarity based smoothing filtering and nonlinear enhancement are combined to remove halftone patterns and preserve the image quality. Experiments demonstrate that the proposed method removes halftone patterns effectively, while preserving more details and recovering cleaner smoothing regions.

**Keywords:** Scanned image, descreening, halftone, adaptive parameters.

## 1 Introduction

Currently, most electrophotographic (EP) printers adopt a halftone technique to approximate the original contone image with a binary halftone image. However, annoying halftone patterns often appear in scanned images of such printed hard copies. These patterns decrease the aesthetic quality of scanned images. Moreover, the periodic halftone patterns introduced by clustered dots halftoning[1], the most commonly used halftoning technique in current EP printers, may produce Moiré effect in hard copies if the scanned images are reprinted[2].

Several methods[3]-[6] recover contone images with details and sharp edges from binary halftone images. Nevertheless, these methods can only work on binary halftone images and they are not suitable to descreen the scanned halftone images because scanned images are grayscale. Some other descreening methods are designed for the scanned halftone image. Intuitively, the simplest way of descreening is to perform low-pass filtering on the scanned image. However, these

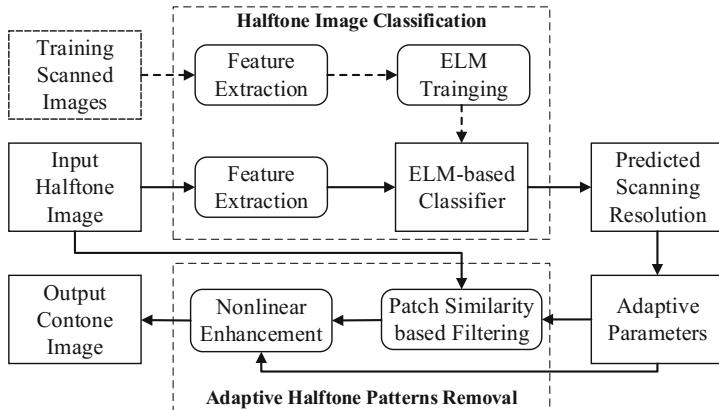
filtering approaches have difficulties in striking a balance between detail preservation and halftone patterns removal. Siddiqui *et al.*[2][7] introduced two descreening methods for scanned halftone images. The training based method[2] contains two steps: basic prediction of contone image, and modified SUSAN filtering based on the predicted version. The authors adopt two schemes to predict the basic contone image, one is simply the Gaussian filtering; the other is resolution synthesis based denoising. And the basic image is used to guide the modified SUSAN filtering to obtain the final descreened image. The results of the method have sharp edges. However, it cannot recover high quality smooth regions. And in the resolution synthesis based denoising, different resolution sample image pairs need to be collected and registered for training. The second method used local gradient information to estimate contone value[7]. Although it has high computational efficiency, this method cannot remove the halftone patterns along edges effectively.

Most of the above methods for scanned images did not pay much attention to the printing and scanning process. In fact, the halftone patterns in the scanned images vary a lot at different scanning resolutions, but most of these methods are not able to select parameters adaptively. This may lead to detail loss and blurred edges in the recovered contone images at some resolutions.

In this paper, a two-step method is proposed to descreen the scanned image adaptively. In the first step, a halftone image classification is proposed to classify the scanned images into different scanning resolutions. The Local Binary Pattern (LBP) feature is extracted and the Extreme Learning Machine (ELM) is used for classification. And in the second step, contone images with high quality are recovered by an adaptive halftone pattern removal algorithm, whose parameters are determined by the classification results. A patch similarity based smoothing filter is used to remove the halftone patterns, and a nonlinear enhancement is followed to improve the details of the recovered contone images.

## 2 Adaptive Scanned Image Descreening

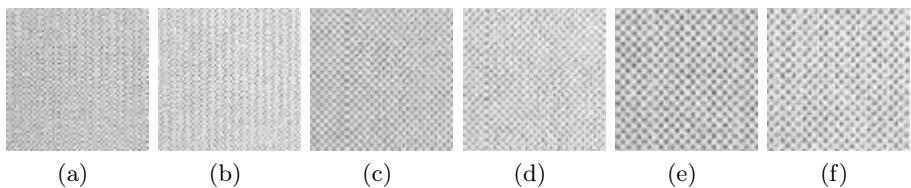
Fig.1 is an overview of the proposed descreening method. It consists of two steps: scanning resolution classification, and adaptive halftone pattern removal. The classification step focuses on distinguishing different scanning resolutions of scanned halftone images. Scanned images with known resolutions make a training dataset. The LBP features of all the scanned images in the training dataset is extracted to train an ELM classifier. For the input scanned image with arbitrary resolution, the ELM classifier is used to predict its scanning resolution. In the adaptive halftone pattern removal step, a patch similarity based smoothing filtering is used to remove halftone patterns of the scanned image, and a nonlinear enhancement is followed to improve the sharpness and contrast of the recovered contone image. The parameters of the patch similarity based smoothing filter and the nonlinear enhancement are adaptively selected by the predicted resolution of the input image.



**Fig. 1.** Overview of the proposed descreening method

## 2.1 Halftone Image Classification

As shown in Fig.2, the forms of halftone patterns vary with the scanning resolutions, but are seldom affected by the different printing setups because the lines per inch(lpi) is a constant in the same printer. In order to achieve better descreening performance, an ELM based classification strategy is proposed to learning the scanning resolution of the scanned image in the first step.



**Fig. 2.** Halftone patterns printed and scanned at different resolutions. (a) Printed at 300 dpi and scanned at 200 dpi. (b) Printed at 600 dpi and scanned at 200 dpi. (c) Printed at 300 dpi and scanned at 300 dpi. (d) Printed at 600 dpi and scanned at 300 dpi. (e) Printed at 300 dpi and scanned at 400 dpi. (f) Printed at 600 dpi and scanned at 400 dpi.

### a) LBP Feature Extraction

The LBP operator[8] has been proved to be an effective texture descriptor. And a widely used extension to the original operator is the so-called uniform patterns represented by  $LBP_{P,R}^{u2}$ , where  $P$  is the number of neighbors and  $R$  is the radius of the neighborhood. Its definition is the binary pattern which contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular. For example, the patterns 00000000, 01000000, and 00111000 are

uniform patterns, while 00101000, 10111000 are non-uniform patterns. In practice, we use uniform patterns with neighborhoods of  $P = 8, R = 1$  computed in the whole scanned image so that the LBP feature histogram has 59 bins.

### b) Classification Using ELM

For classification, ELMs[9] are selected to train the classifier because of its user-friendly and efficient. Each training example is represented by  $(x_i, y_i)$ , where  $i$  is the serial number of training images,  $x_i$  is a 59-dimensional feature vector of scanned image, and  $y_i$  is the associated scanning resolution, for three-class, 200dpi, 300dpi and 400dpi respectively. Sigmoid function is selected as the activation function of ELM. Through tuning the numbers of hidden nodes, we can train the ELM to get the optimal network parameters. And then these parameters are used to make up a classifier for scanned images classification. The related training and testing results are shown in the Section 3.

## 2.2 Adaptive Halftone Pattern Removal

### a) Adaptive Patch Similarity Based Smoothing Filtering

In the proposed descreening algorithm, an adaptive patch similarity based filtering strategy is used to remove halftone patterns in the scanned image. The adaptive patch similarity based filter can be written as follows:

$$v(i) = \frac{1}{z(i)} \sum_{\theta \in \Theta} \sum_{r=1}^R w(i, i_{\theta,r}) h(i_{\theta,r}) \quad (1)$$

where the normalization factor  $z(i)$  is described in (2),  $w(i, i_{\theta,r})$  is the Gaussian kernel which is calculated with (3),

$$z(i) = \sum_{\theta \in \Theta} \sum_{r=1}^R w(i, i_{\theta,r}) \quad (2)$$

$$w(i, i_{\theta,r}) = \frac{1}{z_w} \exp \left( -\frac{d^{patch}(i, i_{\theta,r})}{2\sigma^2} \right) \quad (3)$$

where  $z_w$  is a normalization factor,  $\sigma$  is a scale factor which is determined by the predicted resolution,  $d^{patch}(i, i_{\theta,r})$  measures the similarity of two patches centered at pixel  $i$  and  $i_{\theta,r}$  respectively,  $\theta$  is the directions of patch searching,  $r$  is the searching steps,  $R$  is the searching radius.

To increase the computing efficiency, the patches along the Minimal Hop Path(MHP)[10] are searched in this step. The MHP is the path with the minimal number of hops connecting two patches, and we only consider MHPs along 8 connectivity discrete directions ( $\Theta = \{\frac{i\pi}{4} | i = 1, 2, \dots, 8\}$ ). Considering one MHP direction,  $d^{patch}(i, i_{\theta,r})$  is shown as Eq. (4).

$$d^{patch}(i, i_{\theta,r}) \approx \sum_{t=1}^r \|N(x_t) - N(x_{t-1})\| \quad (4)$$

where  $\theta \in \Theta$ ,  $N(x)$  is the patch centered at pixel  $x$ ,  $r$  is the hop numbers. Eq. (4) can further equal to

$$d^{patch}(i, i_{\theta,r}) \approx d^{patch}(i, i_{\theta,r_1}) + \|N(x_r) - N(x_{r-1})\| \quad (5)$$

It indicates the  $d^{patch}(i, i_{\theta,r})$  can be computed progressively: we can first compute 1-hop path distance and then propagate it to 2-hop, 3-hop, and so on. In practice, we use patch size  $7 \times 7$  and window size  $13 \times 13$ , because smaller size cannot involve enough similarity information, and too large size will decrease the efficiency.

### b) Nonlinear Enhancement

Although the adaptive patch similarity based filtering can remove the halftone patterns effectively, it cannot reduce nonlinear effects introduced in the printing and scanning process, such as contrast reduction. Therefore, a nonlinear enhancement is proposed to improve the contrast and sharpen the edges.

Let the output  $v$  of adaptive patch similarity based filtering be an input image, a Gaussian blurring filter is first performed on  $v$  to get the basic base layer  $v_b$  of  $v$ , shown as Eq. (6):

$$v_b(x, y) = \sum_{i,j} G(i, j)v(x+i, y+j) \quad (6)$$

To avoid strong edges being blurred in the decomposition process, the basic base layer  $v_b$  is refined by the guided filtering[11] with the original image  $v$  serving as the joint image:

$$\hat{v}_b = \text{guidedfilter}(v_b, v) \quad (7)$$

And the detail layer  $v_d$  is obtained by the following equation:

$$v_d = v - \hat{v}_b \quad (8)$$

In order to enhance the details of the recovered contone image, a nonlinear enhancement is performed on the detail layer  $v_d$ . Eq. 9 formalizes the process:

$$\hat{v}_d = HP(s \times Th(v_d)) \quad (9)$$

where  $\hat{v}_d$  is the enhanced detail layer,  $HP$  is the high-pass filtering process,  $s$  is a scaling constant and  $Th(x)$  is the following nonlinear function:

$$Th(x) = \begin{cases} c \cdot x_{max}, & \text{if } x > c \cdot x_{max} \\ x, & \text{if } c \cdot x_{max} \geq x \geq -c \cdot x_{max} \\ -c \cdot x_{max}, & \text{if } x < -c \cdot x_{max} \end{cases} \quad (10)$$

where  $c$  is the clipping constant ranging between 0 and 1. With the enhanced detail layer  $\hat{v}_d$ , the output contone image can be obtained as follows:

$$\hat{I}_o = \hat{v}_b + \hat{v}_d \quad (11)$$

Finally, Eq. 12 is used to improve the image contrast to get the final recovered contone image  $I_o$  of our descreening method.

$$I_o = \frac{\hat{I}_o - \min(\hat{I}_o)}{\max(\hat{I}_o) - \min(\hat{I}_o)} \quad (12)$$

In this enhancement process, the values of  $c$  and  $s$  are adaptively selected in an experiential parameter table by the classification results, and the low-pass Gaussian filter kernel  $G$  and the high-pass filter kernel  $HP$  are given by

$$G = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} / 256, \quad HP = \begin{bmatrix} 1 & -4 & 6 & -4 & 1 \\ -4 & 16 & -24 & 16 & -4 \\ 6 & -24 & 36 & -24 & 6 \\ -4 & 16 & -24 & 16 & -4 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix} / 256.$$

### 3 Experimental Results

In order to evaluate performance of the proposed method on real scanned images, we print and scan the 60 images from the Berkeley Segmentation Database. Each original contone image is printed by printer RICOH Aficio MP4500 at 300 and 600 dpi respectively. And every hard copy is scanned by scanner Fujitsu fi-6130 at 200, 300 and 400 dpi respectively. So for each scanning resolution, 120 images are used to test the proposed method.

#### 3.1 Results of Halftone Image Classification

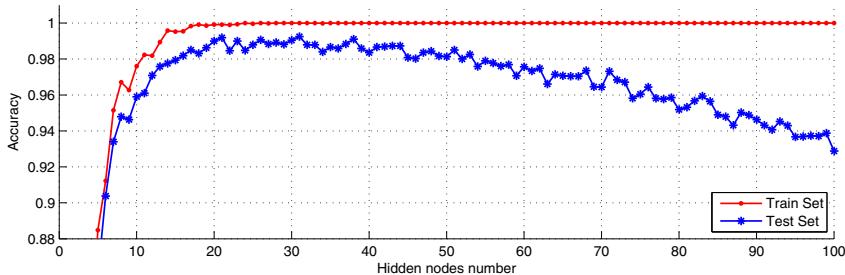
To evaluate the proposed halftone image classification algorithm, the whole dataset listed in Table 1 is equally and randomly partitioned into the training set and test set. Fig. 3 shows the classification accuracies with different numbers of hidden nodes. The result is the average of 20 random experiments. From Fig. 3, it can be found that the accuracy on the training set increases and reaches 100% when the number of hidden nodes is more than 25. The accuracy on the test set increases to 99.4% with about 31 hidden nodes, and decreases gradually if the number of hidden nodes are more than 50, due to the ELM classifier overfitting the training set. In our experiment, the number of hidden nodes is selected to 31.

#### 3.2 Comparison with Existing Methods

The proposed descreening method is compared with three existing methods: training-based descreening using Gaussian filter(TBD-I); training-based descreening using resolution synthesis(TBD-II) and hardware friendly descreening(HFD). The software of the training-based descreening is available on the website. The HFD method is implemented in Matlab. Table 2 shows the parameters of TBD-I, TBD-II and HFD used for the experiments on images scanned at 300 dpi.

**Table 1.** Experimental dataset

Class	Numbers of image	Train Set	Test Set
Original Contone Image	120	60	60
Scanned Halftone Image	Scanned at 200 dpi	120	60
	Scanned at 300 dpi	120	60
	Scanned at 400 dpi	120	60

**Fig. 3.** Accuracy of halftone image classification**Table 2.** Parameter settings of the existing methods used in the experiments

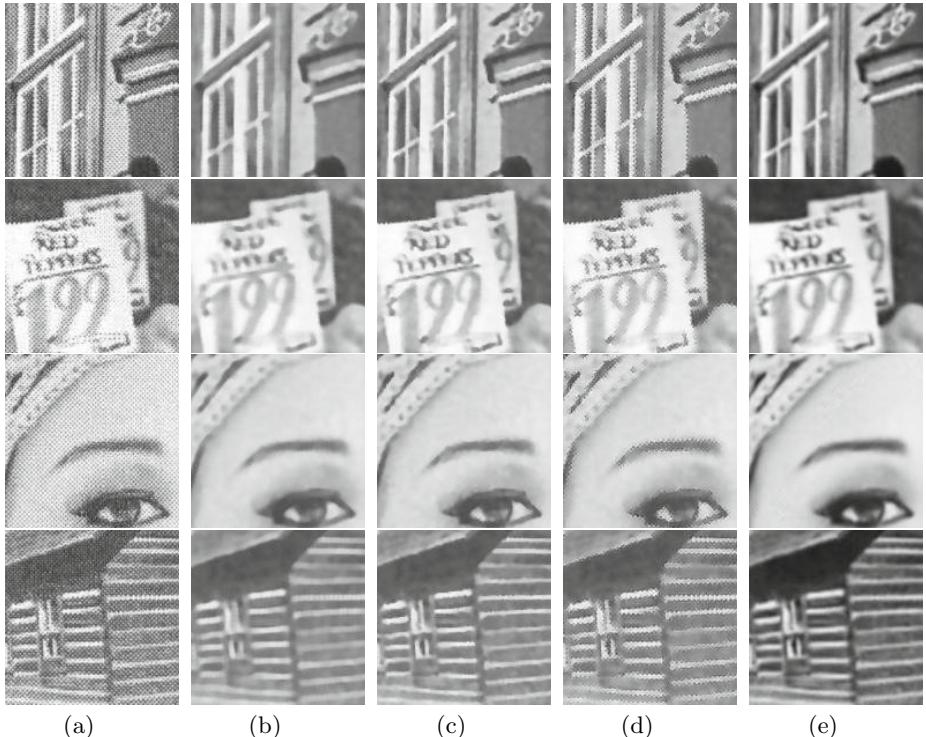
Method	Parameter	Value
TBD-I	Radius for Gaussian filter	3
	Scale factor for Gaussian filter	2.5
	Radius for modified SUSAN filter	3
	Space scale factor	2.5
	Brightness scale factor	21
TBD-II	Delta	2.2
	Radius for modified SUSAN	3
	Space scale factor	2.5
	Brightness scale factor	21
HFD	Filter radius	3
	Sharpness level	0

The parameter values of these three methods are selected as reported in the paper. And the parameter values of the proposed method are listed in Table 3.

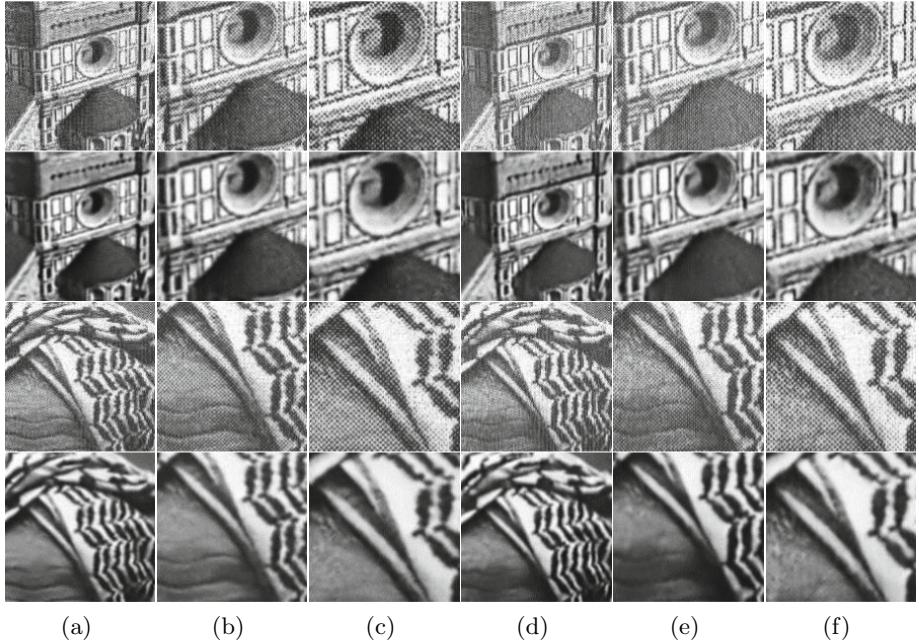
Results of different descreening methods are compared in Fig. 4. Fig. 4(a) is the original scanned image. The descreening results of TBD-I, TBD-II, HFD and the proposed algorithm are shown from Fig. 4(b) to Fig. 4(e) respectively. From Figs. 4(b) and 4(c), the TBD method removes the halftone patterns very

**Table 3.** Parameter settings of the proposed methods used in the experiments

Step	Parameter	Scanned at 200 dpi	Scanned at 300 dpi	Scanned at 400 dpi
Halftone Image Classification	Hidden nodes	31	31	31
Halftone Pattern Removal	Scale factor $\sigma$	22	26	35
	Clipped constant $c$	0.1	0.2	0.3
	Scaling constant $s$	1	2	3

**Fig. 4.** Close-up views of results of different descreening methods. (a) Original scanned images printed at 300 dpi and scanned at 300 dpi. (b) Results of TBD-I. (c) Results of TBD-II. (d) Results of HFD. (e) Results of the proposed method.

well. But the edges produced by the TBD-I method are blurred as shown in Fig. 4(b). The TBD-II method preserves sharp edges, but however, the smooth regions of the results are noisy. In Fig. 4(d), the HFD method cannot remove halftone patterns along the edges. As shown in Fig. 4(e), it can be found that the proposed method removes halftone patterns in smooth regions as well as along edges with improving the image contrast. The recovered edges of the proposed method are clearer and sharper than that of TBD-I method. Compared with



**Fig. 5.** Results of the proposed method on scanned image with different printing and scanning resolutions. (a) The original scanned images printed at 300dpi and scanned at 200dpi and corresponding descreened images. (b) Printed at 300 dpi and scanned at 300 dpi and corresponding descreened images. (c) Printed at 300dpi and scanned at 400 dpi (d) Printed at 600 dpi and scanned at 200 dpi. (e) Printed at 600 dpi and scanned at 300 dpi. (f) Printed at 600 dpi and scanned at 400 dpi.

TBD-II method, the proposed method produces much cleaner smoothing regions with better visual effect.

### 3.3 Different Printing and Scanning Setup

Fig. 5 shows results of the proposed method on scanned images with different printing and scanning resolutions. These results are produced using the same parameters shown in Table 3. From Fig. 5, we can see the proposed method produces very good results for different printing and scanning resolutions. Although halftone patterns become more prominent with the increasing scanning resolution, the details of our result can be still preserved well with only a little decreasing.

## 4 Conclusions

In this paper, a two-step descreening method is proposed to recover high quality contone image for the scanned halftone image. In the first step, an ELM classifier

is used to learn the scanning resolution of the input image by using the LBP features. The predicted resolution is then used for the adaptive parameter selection of the next step. In the second step, the adaptive halftone pattern removal is performed by combining a patch similarity based smoothing filter and a nonlinear enhancement. Experiments on real scanned images show that the proposed descreening method can recover high quality contone images with both sharp edges and clean smooth regions.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant (No. 61172161), the National Natural Science Foundation for Distinguished Young Scholars of China under Grant (No. 61325007).

## References

1. Sullivan, J., Ray, L., Miller, R.L.: Design of Minimum Visual Modulation Halftone Patterns. *IEEE Trans. Syst., Man, Cybern.* 21(1), 33–38 (1991)
2. Siddiqui, H., Bouman, C.A.: Training-Based Descreening. *IEEE Trans. Image Process.* 16(3), 789–802 (2007)
3. Stevenson, R.L.: Inverse halftoning via MAP estimation. *IEEE Trans. Image Process.* 4(4), 486–498 (1997)
4. Chang, P., Yu, C., Lee, T.: Hybrid LMS-MMSE Inverse Halftoning Technique. *IEEE Trans. Image Process.* 10(1), 95–103 (2001)
5. Chen, L., Hang, H.: An Adaptive Inverse Halftoning Algorithm. *IEEE Trans. Image Process.* 6(8), 1202–1209 (1997)
6. Liu, Y., Guo, J., Lee, J.: Inverse Halftoning Based on the Bayesian Theorem. *IEEE Trans. Image Process.* 20(4), 1077–1084 (2011)
7. Siddiqui, H., Bouman, C.A.: Hardware-Friendly Descreening. *IEEE Trans. Image Process.* 19(3), 746–757 (2010)
8. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(7), 971–987 (2002)
9. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501 (2006)
10. Chen, X., Kang, S.B., Yang, J., Yu, J.: Fast Patch-Based Denoising Using Approximated Patch Geodesic Paths. In: Proc. IEEE Conf. Comput. Vision Pattern Recog., pp. 1211–1218 (2013)
11. He, K., Sun, J., Tang, X.: Guided Image Filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part I. LNCS*, vol. 6311, pp. 1–14. Springer, Heidelberg (2010)

# Compressive Sensing Multi-focus Image Fusion

Fang Cheng, Bin Yang<sup>\*</sup>, and Zhiwei Huang

College of Electric Engineering, University of South China, Hengyang, 421001, China  
{chengfang1990110, yangbin01420}@163.com,  
fuzhi619@aliyun.com

**Abstract.** Based on the compressive sensing theory (CS), various compressive imaging (CI) systems have been developed. Meanwhile, image fusion methods that directly perform on the measurements from multiple CI sensors are also investigated in literatures. In this paper, we presented a multi-focus image fusion method in compressive sensing domain. The main contribution is to introduce a novel clarity level of the random CI measurements without prior geometric information. The CI measurements are sparsely represented with DCT bases which are also projected into the CS domain. Then the sparse coefficients responding to DCT bases are used to guide the fusion of CI measurements of CI sensors. Finally, the fused images are obtained with CS recovery algorithm based on the block compressive sensing (BCS) theory. The simulation results validate the proposed method.

**Keywords:** image fusion, compressive sensing, compressive imaging, focus measure.

## 1 Introduction

Multi-focus image fusion, falling into the category of image fusion, aims at creating a synthetic “all-in-focus” image from input images with a finite depth of field [1]. It is broadly used in many applications such as remote sensing, target identification, medical imaging and so on [2]. In the past two decades, various multi-focus image fusion algorithms have been developed, and these methods can be classified roughly into spatial domain-based and transform domain-based methods. An intuitive spatial domain-based fusion approach is to average the inputs by using different weighting schemes. The transform domain-based methods usually include three steps briefly. Firstly, each input image is decomposed into multi-scale multi-orientation coefficients. Then the transformed coefficients are fused with some fusion rules accordingly. Finally, the fused image is reconstructed by implementing inverse transforms. The multi-scale transform such as the Laplace pyramid [3], gradient pyramid [4], wavelet transform [5], discrete cosine transform (DCT) [6] and nonsubsampled contourlet transform (NSCT) [7] are widely used to perform the fusion.

---

<sup>\*</sup> Corresponding author.

For both spatial and transform domain-based methods, the full resolution source images have to be acquired previously, which increases the storage burden due to the growing sensor data volumes. Fortunately, a new signal acquisition technique known as compressive sensing (CS) [8] has been developed, which accurately reconstructs the original signal from a relatively small number of linear, non-adaptive measurements only if the signal is sparse or compressive in some orthogonal bases. Based on the principle of CS theory, various compressive imaging (CI) techniques have been developed, meanwhile different image fusion methods for CI measurements are also presented [9][10][11]. The scheme adopted in [9] realizes the fusion by taking a “double-star partial Fourier matrix” sampling pattern and simply fusing the measurements with maximum absolute fusion rule. However, the partial Fourier matrix is only incoherent with sparse signals in time domain which restricts its practical applications. The authors in [10] represented an image fusion scheme with DCT sampling model. The CI measurements of multiple source images are combined by weighting their wavelet coefficients. Luo et al [11] introduced a fusion algorithm with scrambled block Hadamard ensemble, a fast measurement operator for CI. The measurements are fused with a weighted average rule based on the entropy of measurements. However, it suffers from some undesirable side effects when the entropy of one measurement fails to measure the activity level. Obviously, exploring the clarity level of each measurement is the key of multi-focus image fusion in compressive sensing domain.

According to CS theory, the CI measurements hold the most information and energy of the scene [12]. The key problem is how to measure the salient information or activity levels of those CI measurements. Inspired by the idea that DCT coefficients can be adopted to measure the activity level of source images [13][14]. We expect to project the CI measurements into the DCT domain. To achieve this aim, the CI measurements are sparsely represented with DCT bases which are also projected into the CS domain. Thus the sparse coefficients responding to a DCT basis are used to guide the fusion of CI measurements with different CI sensors. The activity levels of CI measurements are calculated from the sparse coefficients. Higher activity level means more salient information contained in the relevant CI measurement. Based on the above analysis, after comparing the new DCT coefficients of CI measurement vectors of multiple input images, we select the ideal CI measurement accordingly. At last, using BCS theory we can get final fused image as expected. Simulation results show that the proposed fusion scheme achieves better performances in both subjective and objective qualities than the other fusion schemes.

The remainder of the paper is structured as follows: In section 2, clarity level of CI measurement is presented. In section 3, the design of fusion rule in compressive sensing domain and the schematic diagram of the proposed fusion scheme are given. Simulation results of four pair multi-focus images are presented in Section 4 and finally Section 5 concludes this paper.

## 2 Clarity Level of CI Measurements

The CS theory guarantees accurate reconstruction of the original signal from a relatively small number of linear, non-adaptive measurements by solving an under-determined system

$$\min \|\mathbf{Wx}\|_0 \text{ s.t. } \mathbf{y} = \Phi \mathbf{x}, \quad (1)$$

where  $\mathbf{y} \in \mathbb{R}^M$  is the measurement vector,  $\mathbf{x} \in \mathbb{R}^N$  is the compressible original signal. The measurement matrix  $\Phi$  is a random matrix with a size of  $M \times N (M \ll N)$ . Therefore Eq. (1) is an under-determined problem. The regularization factor  $\|\mathbf{Wx}\|_0$  counts the number of non-zero elements of  $\mathbf{Wx}$ . The matrix  $\mathbf{W}$  is constructed with wavelet bases, which transform original signal  $\mathbf{x}$  into its sparse coefficients. In order to reduce the size of measurement matrix  $\Phi$ , block compressive sensing (BCS) is developed. The difference with the traditional CS system is that instead of sampling the whole image using a fairly large measurement operator, BCS divides the original image into small blocks firstly and samples of each block independently sensed using a comparable small measurement matrix[15][16].

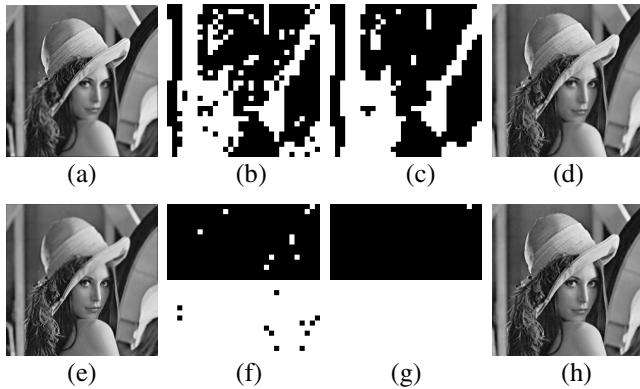
According to the CS theory, various CI systems have been developed. Compared to conventional optical imaging system based on Nyquist theorem, CI system can significantly reduce the computational complexity and storage load at acquisition stage. One of the most famous applications of CI is single-pixel camera developed by Rice University [17]. The novel digital image camera directly acquires random projections of a scene without collecting the pixels. Its key hallmark is the ability to obtain an image with a single detection element while measuring the scene fewer times than the total number of pixels of the image.

Suppose  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are the CS measurement vectors of different CI cameras. The key problem for measurements fusion is the actively level evaluation of  $\mathbf{y}_1$  and  $\mathbf{y}_2$  since they contain no geometry information. As indicated in [14], DCT coefficients of image block have been effectively used to calculate the clarity level of multi-focus images. However, we can not obtain the DCT coefficients of  $\mathbf{y}$  in CS domain directly. Suppose  $\mathbf{x}$  is the original 1D image block of the CS measurement  $\mathbf{y}$ , then

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} \quad (2)$$

where  $\mathbf{D}$  is the DCT bases, vector  $\boldsymbol{\alpha}$  is the DCT coefficient of  $\mathbf{x}$ . When we transform both the original 1D image block  $\mathbf{x}$  and  $\mathbf{D}$  into the CS domain as  $\mathbf{y}$  and  $\hat{\mathbf{D}} = \Phi \mathbf{D}$  respectively, the DCT coefficient  $\boldsymbol{\alpha}$  would not be changed according to the restricted isometry property [18] of  $\Phi$ . Therefore, the CS measurement  $\mathbf{y}$  can be represented as  $\mathbf{y} = \hat{\mathbf{D}}\boldsymbol{\alpha}$ . Since the dimension of  $\boldsymbol{\alpha}$  is larger than that of  $\mathbf{y}$ , we can not calculate  $\boldsymbol{\alpha}$  directly. In this paper, we use the orthogonal matching pursuit (OMP) [19] to obtain the approximate DCT coefficient. The  $\ell_1$ -norm of  $\boldsymbol{\alpha}$  is used to represent the clarity level of the CS measurement  $\mathbf{y}$ .

Without loss of generality, we take Lenna image with a size of  $256 \times 256$  as example. By blurring the upper and lower parts respectively, we can get a pair of artificial multi-focus images as in Fig.1 (a) and Fig.1 (e). The CS measurements of each  $8 \times 8$  block of the two images are obtained by the same random CS measurement matrix. Then the clarity levels of the CI measurements are calculated with the proposed method and the absolute of measurements respectively. Comparing the clarity level of the CI measurements, the clarity maps which indicate the clarity distribution are obtained as shown in Fig.1 (b)(c) and Fig.1 (f)(g) respectively. Fig.1 (d) and (h) gives the fused image with the max-abs and proposed scheme respectively. The example demonstrates the validation of the proposed clarity level.



**Fig. 1.** An example of artificial multi-focus images to demonstrates the validation of the proposed clarity level. (a) and (e) artificial multi-focus Lenna images; (b) and (f) clarity maps of Max-abs and proposed scheme respectively; (c) and (g) clarity maps after majority filter; (d) and (f) the fused images of two fusion schemes.

### 3 The Proposed Fusion Framework

Fig.2 illustrates the schematic diagram of the proposed fusion scheme in CS domain. We assume that there are two CI cameras for simplicity. With different focus settings, the CI measurement vectors have finite depth of field. Our aim is to combine the CI measurement vectors of different CI cameras into an integrated CI measurement vectors in the CS domain. The proposed fusion scheme includes three major steps:

(1) The real scene is sensed by two different CI cameras based on the BCS theory. With different focus settings, different optical images are sensed by the digital micromirror device of the CI cameras. We assume the two optical images are refers as  $\mathbf{I}_1$  and  $\mathbf{I}_2$  respectively. Let  $\Phi_B$  denotes the orthonormal measurement random Gaussian matrix. The CI measurement vectors can be represented as

$$\mathbf{y}_1^i = \Phi_B \mathbf{R}_i \mathbf{I}_1 \quad (3)$$

$$\mathbf{y}_2^i = \Phi_B \mathbf{R}_i \mathbf{I}_2 \quad (4)$$

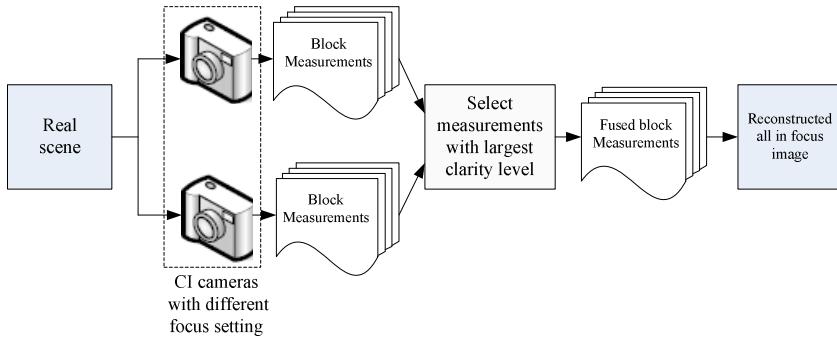
where  $\mathbf{R}_i$  extracts the  $i$ th block of optical image  $\mathbf{I}_1$  and  $\mathbf{I}_2$ . The vectors  $\mathbf{y}_1^i$  and  $\mathbf{y}_2^i$  are the CI measurements obtained by CI cameras.

(2) Since  $\mathbf{y}_1^i$  and  $\mathbf{y}_2^i$  response to the different focus settings measurements of the same scene, we need to fuse them into a composite vector  $\mathbf{y}_F^i$  which is the measurement of the fused all-in-focus image. We select the measurements with a largest clarity level directly as

$$\mathbf{y}_F^i = \begin{cases} \mathbf{y}_1^i & C_1^i \geq C_2^i \\ \mathbf{y}_2^i & \text{elsewise} \end{cases} \quad (5)$$

where  $C_1^i$  and  $C_2^i$  are the clarity level of  $\mathbf{y}_1^i$  and  $\mathbf{y}_2^i$ .

(3) The synthetic all-in-focus image from the integrated CI measurement  $\mathbf{y}_F^i$  with the basic ingrained directional transforms SPL-based CS reconstruction framework was proposed in [20]. Among all the directional transforms, we choose dual-tree discrete wavelet transform (DDWT) due to easy implementation.



**Fig. 2.** Framework of the proposed fusion method

## 4 Simulation Results

In this section, we evaluate the performance of the proposed fusion scheme through computer simulation. Four pairs natural images (shown in Fig. 3) which are common adopted in literatures are employed as the optical images with different focus settings of CI sensors. The CI measurements are simulated by Eq. (3) and (4) where the measurements matrix  $\Phi_B$  are generated by a random Gaussian process. The global error parameter  $\epsilon$  is set to be a small value for avoiding losing detail information of the measurements. We stressed that this parameter can be tuned according to the noise level when considering the effectiveness of noise. The DCT dictionary is obtained by sampling the cosine wave at different frequencies. Two objective evaluation criteria,

$Q_w$  [21] and  $Q^{AB/F}$  metric [22], are used to perform the subjective evaluation. Both measures should be as close to 1 as possible.



Fig. 3. Four pair of natural test source images

To demonstrate the effectiveness of the proposed fusion scheme, we compare it with the absolute maximum select rule [9] and the linearly weighted average based on entropy [10]. Different block sizes lead to different performances. In general, under a fixed sampling rate, the bigger block size is set, the better performance will be achieved. Without loss of generality, three sampling rates 0.3, 0.5, 0.7 are used in this paper. At each sampling rate, each pair of source images with a block size of 8, 16, and 32 are set for BCS. In order to explore the effect of block size on the results, we take Clock image as an example. The objective evaluation criteria  $Q_w$  and  $Q^{AB/F}$  of the fused images are indicated in Fig.5 and Fig.6.

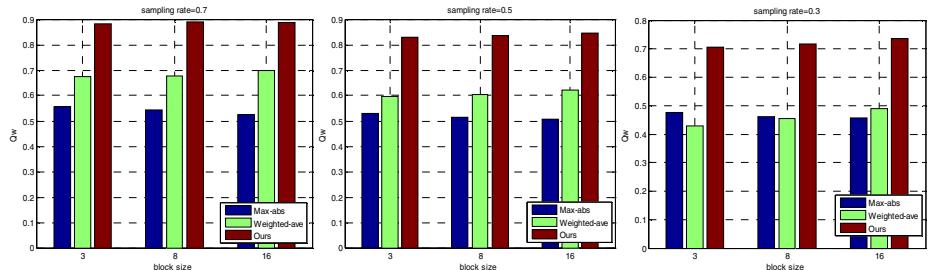
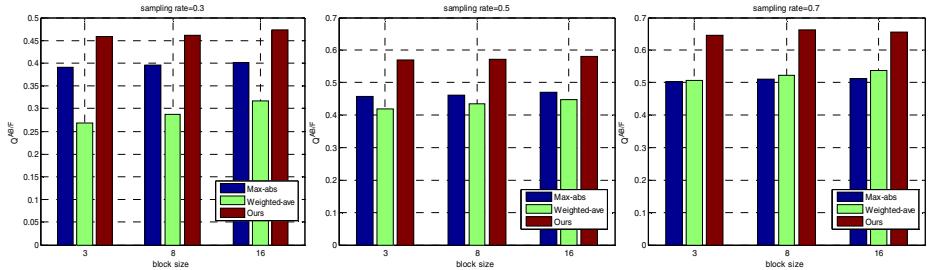


Fig. 4.  $Q_w$  of different sampling rates with different block sizes

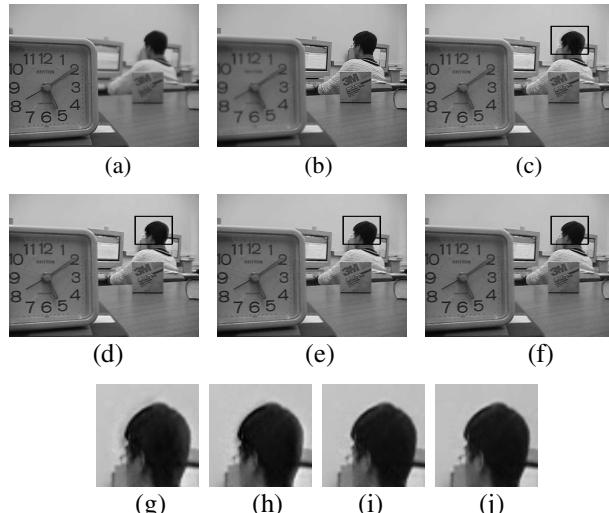
In Fig.5 and Fig. 6, the proposed fusion scheme shows significant advantages over other two kinds of methods; and better performance will be achieved as the block size increases for the both two metrics. In the following experiments, the block size is set to 16, considering not only performances, but also calculation time. Fig. 7 illustrates the fused results of three methods for image Clock.



**Fig. 5.** QAB/F of different sampling rates with different block sizes



**Fig. 6.** Fusion results for Clock image. (a) fused image with Max-abs rule; (b) fused image of Weighted-ave; (c) fused image of our proposed fusion scheme.



**Fig. 7.** Fusion results for Lab image. (a) and (b) original source images; (c) fused image of DTCWT; (d) fused image of NSCT; (e) fused image of GFF; (f) fused image of our proposed fusion scheme; (g)(h)(i)(j) are the local magnified version of (c)(d)(e) and (f).

In addition, the proposed method is compared with existing multifocus image fusion methods involving the dual-tree complex wavelet transform (DTCWT) [23], the nonsubsampling contourlet transform (NSCT) [7] and the guide filter (GFF) [24] based methods. For those methods, the sensed images are reconstructed by BCS from

the measurements firstly. Then the reconstructed images are fused based on various methods. For all DTCWT, and NSCT-based methods, the input images are decomposed with four levels. For NSCT, 2, 8, 8, and 16 directions are used in the high-frequency scales. The absolute value of the coefficients are employed as the activity level and the absolute-max fusion rule with  $3 \times 3$  widow consistency verification fusion scheme are adopted to fuse the transform coefficients.

Without loss of generality, we present results at three sampling rates 0.3, 0.5, and 0.7 respectively. At each sampling rate, we conduct our experiments on each pair of source images. As described above, the proposed fusion rule for CI measurements performs better than the Max-abs and the linearly weighted average of CI measurements. , Because our proposed fusion scheme transfers more edge and detail information from source images to fused image. Figure 8 gives a sample to illustrate the fused images of different methods.

By carefully observing the fusion results, the fused image of DTCWT results in the side effect of reduction in contrast and on the back of the man's head, it suffers from a kind of ringing artifact. In addition, there are some artifacts for the NSCT and GFF. As it can be seen in the magnified images, our proposed algorithm has a considerable improvement in fused image quality. In conclusion, simulation results indicate that the proposed multi-focus images fusion scheme not only outperforms the state-of-the-art fusion rule for CI measurements, but also shows comparable level performance with most multi-scale decomposition fusion methods. Objective evaluation results  $Q_w$  and  $Q^{AB/F}$  of the test images are listed in Table 1.

**Table 1.** Quantitative assessments of various fusion methods

Images	Sampling rates	Methods: DTCWT NSCT GFF CS					
		$Q_w$	$Q^{AB/F}$				
clock	0.3	0.7102 0.7166 0.7103 <b>0.7180</b>	0.4444 0.4551 0.4479 <b>0.4607</b>				
	0.5	0.8275 0.8302 0.8267 <b>0.8366</b>	0.5634 0.5709 0.5659 <b>0.5718</b>				
	0.7	0.8849 0.8863 0.8845 <b>0.8906</b>	0.6511 0.6580 0.6597 <b>0.6622</b>				
pepsi	0.3	0.5607 <b>0.5648</b> 0.5607 0.5619	0.4515 <b>0.4589</b> 0.4485 0.4526				
	0.5	0.7097 0.7121 0.7105 <b>0.7293</b>	0.5411 0.5472 0.5416 <b>0.5510</b>				
	0.7	0.8292 0.8314 0.8289 <b>0.8483</b>	0.6194 0.6242 0.6218 <b>0.6375</b>				
ball	0.3	0.6664 <b>0.6695</b> 0.6610 0.5285	0.5136 0.5192 0.5100 <b>0.6658</b>				
	0.5	0.8049 0.8066 0.8002 <b>0.8104</b>	0.6652 0.6709 0.6625 <b>0.6801</b>				
	0.7	0.8828 0.8842 0.8812 <b>0.8883</b>	0.7597 0.7639 0.7584 <b>0.7704</b>				
lab	0.3	0.7123 0.7174 0.7133 <b>0.7220</b>	0.4148 <b>0.4269</b> 0.4112 0.4261				
	0.5	0.8342 0.8364 0.8348 <b>0.8422</b>	0.5384 0.5463 0.5381 <b>0.5517</b>				
	0.7	0.8848 0.8860 0.8851 <b>0.8890</b>	0.6088 0.6182 0.6147 <b>0.6255</b>				

## 5 Conclusion

This paper proposes a novel image clarity level in the CS domain. Based on the clarity level, a multi-focus image fusion scheme is proposed in the CS domain for CI sensors. Natural multi-focus images are used to validate the proposed method. In comparison with the existing fusion schemes in CS domain and the traditional multi-resolution-based methods, the advantages of the proposed method are confirmed.

**Acknowledgements.** This paper is supported by the National Natural Science Foundation of China (Nos. 61102108, 11247214 and 61172161), Scientific Research Fund of Hunan Provincial Education Department (Nos. 11C1101 and 12A115), and the construct program of key disciplines in USC (No.NHJK04).

## References

- [1] Li, H., Chai, Y., Yin, H., et al.: Multifocus image fusion and denoising scheme based on homogeneity similarity. *Optics Communications* 285(2), 91–100 (2012)
- [2] Qu, G.H., Zhang, D.L., Yan, P.F.: Medical image fusion by wavelet transform modulus maxima. *Optics Express* 9(4), 184–190 (2001)
- [3] Liu, Z., Tsukada, K., Hanasaki, K., et al.: Image fusion by using steerable pyramid. *Pattern Recognition Letters* 22(9), 929–939 (2001)
- [4] Petrovic, V.S., Xydeas, C.S.: Gradient-based multiresolution image fusion. *IEEE Transactions on Image Processing* 13(2), 228–237 (2004)
- [5] Pajares, G., Manuela de la Cruz, J.: A wavelet-based image fusion tutorial. *Pattern Recognition* 37(9), 1855–1872 (2004)
- [6] Chu, H., Zhu, W.L.: Image fusion algorithms using discrete cosine transform. *Optics and Precision Engineering* 14(2), 266–273 (2006)
- [7] Zhang, Q., Guo, B.: Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Processing* 89(7), 1334–1346 (2009)
- [8] Baraniuk, R.: Compressive sensing. *IEEE Signal Processing Magazine* 24(4), 118–121 (2007)
- [9] Wan, T., Canagarajah, N., Achim, A.: Compressive image fusion. In: *Proceedings of 15th IEEE International Conference on Image Processing*, pp. 1308–1311 (2008)
- [10] Han, J., Loffeld, O., Hartmann, K., et al.: Multi image fusion based on compressive sensing. In: *Proceedings of the International Conference on Audio Language and Image Processing*, pp. 1463–1469 (2010)
- [11] Luo, X., Zhang, J., Yang, J.Y., Dai, Q.H.: Image fusion in compressed sensing. In: *Proceedings of 16th IEEE International Conference on Image Processing*, Piscataway, NJ, pp. 2205–2208 (2009)
- [12] Luo, X., Yang, J., Dai, Q., et al.: Classification-based image-fusion framework for compressive imaging. *Journal of Electronic Imaging* 19(3), 033009-1–033009-14 (2010)
- [13] Yang, B., Li, S.T.: Pixel-level image fusion with simultaneous orthogonal matching pursuit. *Information Fusion* 13(1), 10–19 (2012)
- [14] Haghighat, M.B.A., Aghagolzadeh, A., Seyedarabi, H.: Multi-focus image fusion for visual sensor networks in DCT domain. *Computers & Electrical Engineering* 37(5), 789–797 (2011)

- [15] Gan, L.: Block compressed sensing of natural images. In: Proceedings of 15th IEEE International Conference on Digital Signal Processing, pp. 403–406 (2007)
- [16] Mun, S., Fowler, J.E.: Block compressed sensing of images using directional transforms. In: Proceedings of 16th IEEE International Conference on Image Processing, pp. 3021–3024 (2009)
- [17] Baraniuk, R.G.: Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine* 25(2), 83–91 (2008)
- [18] Candès, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, vol. 3, pp. 1433–1452 (2006)
- [19] Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory* 53(12), 4655–4666 (2007)
- [20] Mun, S., Fowler, J.E.: Block compressed sensing of images using directional transforms. In: Proceedings of 16th IEEE International Conference on Image Processing, pp. 3021–3024 (2009)
- [21] Piella, G.: H Heijmans. A new quality metric for image fusion. In: Proceedings of the International Conference on Image Processing, vol. 2, pp. III-173–III-176 (2003)
- [22] Xydeas, C.S., Petrović Objective, V.: image fusion performance measure. *Electronics Letters* 36(4), 308–309 (2000)
- [23] Kingsbury, N.: Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis* 10(3), 234–253 (2001)
- [24] Li, S.T., Kang, X.D., Hu, J.W.: Image fusion with guided filtering. *IEEE Transactions on Image Processing* 22(7), 2864–2875 (2013)

# Pan-Sharpening Based on Improvement of Panchromatic Image to Minimize Spectral Distortion

Akbi Abdelkrim, Zhaoxiang Zhang, and Qingjie Liu

IRIP Lab. School of Computer Science and Engineering, BEIHANG University,  
Beijing, China

**Abstract.** In this paper, we propose a novel method to enhance the pan-sharpening result of low-resolution multispectral images (MS) and high-resolution panchromatic images (Pan) by minimizing the spectral distortion engendered by the fusion process. In fact, spectral distortion is the most significant problem in many pan-sharpening techniques, due to the non-linearity between Pan and MS images. In this method, an improvement of the Pan image is performed in order to enhance the correlation between Pan and MS images before pan-sharpening process. The proposed method is applied as a preprocessing by fusing the intensity image derived from MS image with the original Pan to get an improved Pan image which could be more correlated with MS image. And later, the pan-sharpening is applied on both MS and the improved Pan using any pan-sharpening technique. Simulation results of proposed method are compared in four different techniques, such as: Generalized IHS, DWT, Brovey and HPF. It has been observed that simulation results of this method preserves more spectral information and gets better visual quality compared with earlier reported techniques using original Pan.

**Keywords:** Pan-sharpening, Fusion, Panchromatic, Multispectral, Spectral distortion.

## 1 Introduction

Image fusion has become a very important domain during the last two decades due to the huge acquirement of images such as remote sensing images for earth observation. Satellite sensors offer Pan images with high spatial resolution and low spectral resolution and MS images with low spatial resolution and high spectral resolution on the same observation region. Therefore, both images should be fused (pan-sharpened) to produce images with high spatial and high spectral resolution. Different methods are developed in this sense, but with the limitation of sensors acquisition, Pan and MS images are not well correlated to produce a good pan-sharpened image which should be as close as possible to those that would have been acquired by the corresponding sensors if they had the high resolution of Pan. In fact, this limitation causes the spectral or color distortion on pan-sharpened images which is the most significant problem in pan-sharpening processing. Therefore, Other strategies should be taken into account in order to deal with this issue. Among that, statistical techniques allow to accurately model the relationship between the MS and Pan images [1].

To reduce the color (spectral) distortion and improve the fusion quality, a wide variety of strategies have been developed, each specific to a particular fusion technique or image dataset. No satisfactory solution has been achieved which can consistently produce high quality fusion for different data sets as well as it can reduce color distortion [2]. The reason of color distortion is a criteria that should be considered. Thus, the source of this issue is the limitation of satellite sensors capabilities in the process of image acquirement (spectral response, noise properties...) and the conditions in which the images were taken.

Before research in the field of Pan-sharpening we should have knowledge about the behaviors of satellite sensors and how they produce images either Pan or MS. The issue of spectral distortion is one of the hot topics in the current research of the fusion methods based on the combination models. Some researchers proposed some improved combination models [3] or combined different types of pan-sharpening methods [4] to get better fusion results. Most pan-sharpening techniques use actually histogram matching of the Pan and MS bands as a preprocessing step in order to minimize brightness mismatching during the fusion process, which may help to reduce the spectral distortion in the pan-sharpened image [1].

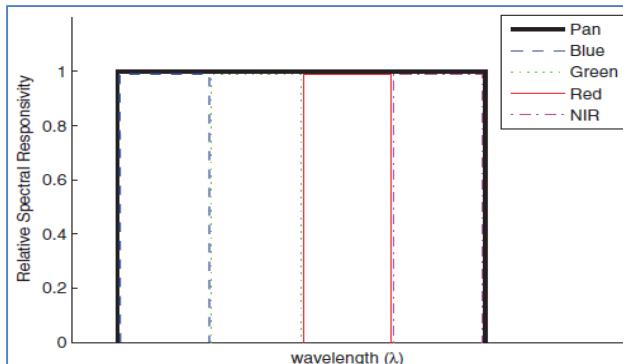
In this paper, we propose a novel preprocessing method for pan-sharpening techniques to further improve their effectiveness. This method is applied to enhance the Pan image, based on the linear model, in which the original Pan is a linear combination with original MS bands. According to this model, an improved Pan image is generated by fusing the intensity image of MS bands with the original Pan. Thus, spectral information can be injected from MS into the new Pan leading to enhance the correlation between them, and later, Pan-sharpening is performed on both improved Pan and MS images to reduce the spectral distortion.

## 2 Linear Model

We begin by formulating an observational model. This model is based on the assumptions that the decimated pan-sharpened image gives the MS image and that the Pan image is a linear combination of the bands of the MS image [5]. If MS and Pan sensors all work at the same spatial resolution, we can assume that the radiance energy of the same ground instantaneous field of view (GIFOV) captured by these two sensors is equal, i.e.

$$E_P = E_{MS} = \sum_{i=1}^N E_i$$

where  $E_P$  and  $E_{MS}$  are energy of Pan and energy of MS.  $N$  is the number of bands of the MS sensors. Under the ideal theoretical condition, the spectral response functions of all these sensors are all the same, and the response curve of Pan sensor covers the curves of MS sensors exactly as shown in Figure 1, the response signal of the Pan sensor is also the sum of MS signals [6].

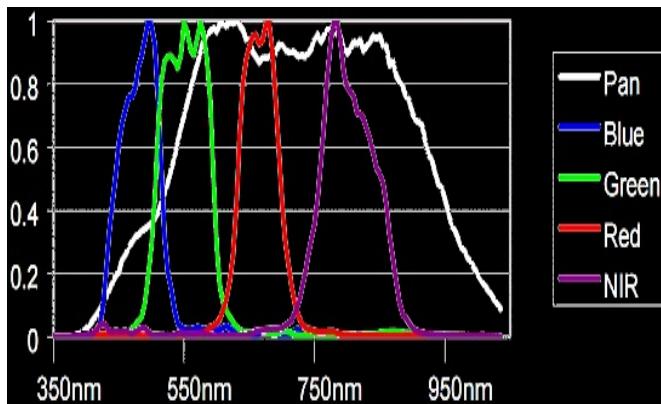


**Fig. 1.** Normalized spectral response curve of Pan and MS sensors for ideal theoretical condition

In the case of real sensors the spectral response functions are much more complicated. For example, Figure 2 shows the IKONOS spectral response curves. There is obvious overlap, missing, and overflow among the curves of Pan and MS bands. The relation between Pan and MS is not exactly linear. So, the equation above can be written as:

$$B_P = \sum_{i=1}^N \alpha_i B_i + c, \quad \alpha_i > 0.$$

Where  $B_P$  and  $B_i$  indicate the pixel value of Pan and the  $i^{\text{th}}$  band of MS,  $\alpha_i$  indicate the weight of the band  $B_i$ , and  $C$  is the real difference between Pan and MS or the missing information to get the ideal model causing the problem of spectral distortion.



**Fig. 2.** IKONOS spectral response curves of Pan and MS bands

### 3 Proposed Method

According to the model that we have seen before, many researchers attempt to minimize the deference between Pan and MS images in order to avoid spectral distortion, due to the non-linearity between them. Lining Liu et al [7] has created a new MS image (MS+) by appending this difference as a new band such as it fits better with the original Pan.

In the proposed method we try first to find an improved Pan image ( $P^*$ ) as the same resolution of the original Pan ( $P$ ) which could be more correlated with the MS image. However, among the objectives of image fusion is to substitute missing information from input images in order to get best informative image [8]. So, The main idea is to generate the improved Pan ( $P^*$ ) by fusing the original Pan ( $P$ ) with the Intensity image ( $I^*$ ) derived from MS. In this case, ( $P^*$ ) might have more spectral information derived from MS than the original Pan ( $P$ ). After that, the pan-sharpening process will be carried out on both improved Pan ( $P^*$ ) and MS images. Thus, the present analysis is to perform two levels of fusion:

- First level (*preprocessing*):

- Generate the intensity Image ( $I^*$ ) from the MS image such as:

$$I^* = \sum_{i=1}^N \alpha_i B_i$$

$\alpha_i$  is the weight of the  $i^{th}$  band  $B_i$  for the MS image with  $N$  bands, in particular, we can choose:

$$\alpha_i = \frac{1}{N}$$

To improve the correlation between MS and Pan images we can determine  $\alpha_i$  by solving this equation using the least square fitting method:

$$A = \left( X^T X \right)^{-1} X^T Y$$

where:

$$A = [\alpha_1, \alpha_2, \dots, \alpha_N]^T,$$

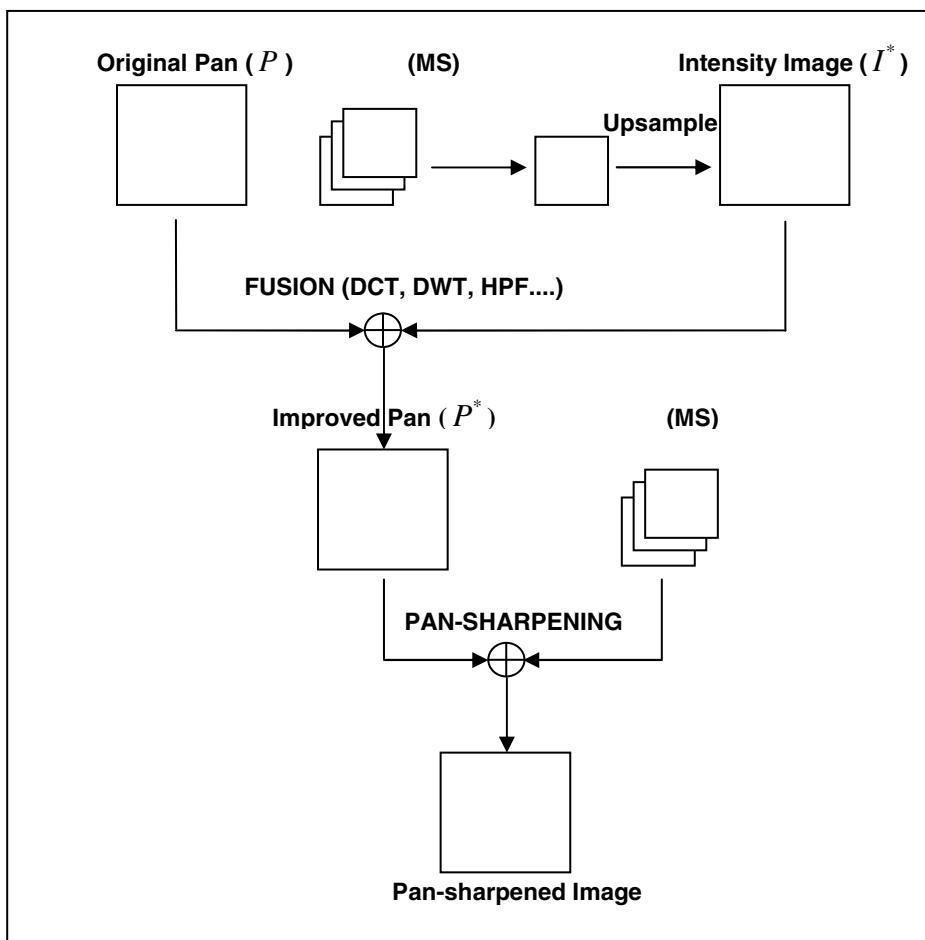
$$Y = [I_1^*, I_2^*, \dots, I_M^*]^T$$

and:

$$X = \begin{bmatrix} B_{11} & \cdots & B_{N1} \\ \vdots & \ddots & \vdots \\ B_{1M} & \cdots & B_{NM} \end{bmatrix}$$

$M$  denotes the number of pixels and  $N$  denotes the number of bands in MS image.

- Upsample the generated intensity image ( $I^*$ ) to the size of the original Pan ( $P$ ).
- In order to preserve better spatial information of the original Pan ( $P$ ), the later is fused with the upsampled Intensity image ( $I^*$ ) using one of the frequency domain fusion techniques (DCT, DWT, HPF...). This domain is known as a powerful tool to extract high frequency information (details) from images. Finally, the result is an improved panchromatic image ( $P^*$ ) which can take a good adjustment of spectral information by injecting them from the original MS.
- Second level (*pan-sharpening process*):
- Pan-sharpen the original MS with the improved Pan ( $P^*$ ) using any pan-sharpening technique. Figure.3 shows the diagram of the proposed method.



**Fig. 3.** Diagram of the proposed method

## 4 Quality Assessment

According to the protocol proposed by Wald et al [9], the pan-sharpened images are downsampled to the size of the original MS images. Hence, the results are assessed both visually and quantitatively using the following indicators:

- Correlation coefficient (*CC*) between the original and the pan-sharpened images. It should be as close as possible to 1 [10].

$$CC(A, B) = \frac{\sum_{m,n} (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_{m,n} (A_{mn} - \bar{A})^2)(\sum_{m,n} (B_{mn} - \bar{B})^2)}}$$

where  $\bar{A}$  and  $\bar{B}$  stand for the mean values of the corresponding data set, and *CC* is calculated globally for the entire image. The result of this equation shows similarity in the small structures.

- The relative average spectral error (*RASE*) index [11]. It characterizes the average performance of the method of image fusion in the spectral bands considered:

$$RASE = \frac{100}{M} \sqrt{\frac{1}{k} \sum_{i=1}^k RMSE^2(B_i)}$$

where *M* is the mean radiance of the *k* spectral bands  $B_i$  of the original MS bands, and *RMSE* is the root mean square error computed in the following expression:

$$RMSE(B_k) = \sqrt{\frac{\sum_{i=1}^N (B_k(i) - B_k^*(i))^2}{N}}$$

where *N* is the number of pixels,

$B_k$  is the original image for the band *k* and

$B_k^*$  is the fused image in band *k*.

- The erreur relative globale adimensionnelle de synthèse (*ERGAS*) index (or relative global dimensional synthesis error) in the fusion [12]:

$$ERGAS = 100 \frac{h}{l} \sqrt{\frac{1}{k} \sum_{i=1}^k \frac{RMSE^2(B_i)}{M_i^2}}$$

where *h* is the resolution of the high spatial resolution image and *l* is the resolution of the low spatial resolution image and  $M_i$  the mean radiance of each spectral band  $B_i$  involved in the fusion. The lower the value of the *RASE* and *ERGAS* indexes, the higher the spectral quality of the fused images.

- Spatial Coefficient (SC): A good fusion method must allow the addition of a high degree of the spatial detail of the Pan image to the MS image. The addition of this spatial detail is evident for all the merged images when these are visually compared to the initial MS. To evaluate this spatial detail addition, we used the procedure proposed by Zhou [13]. The Pan and fused images are filtered using the Laplacian filter:

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

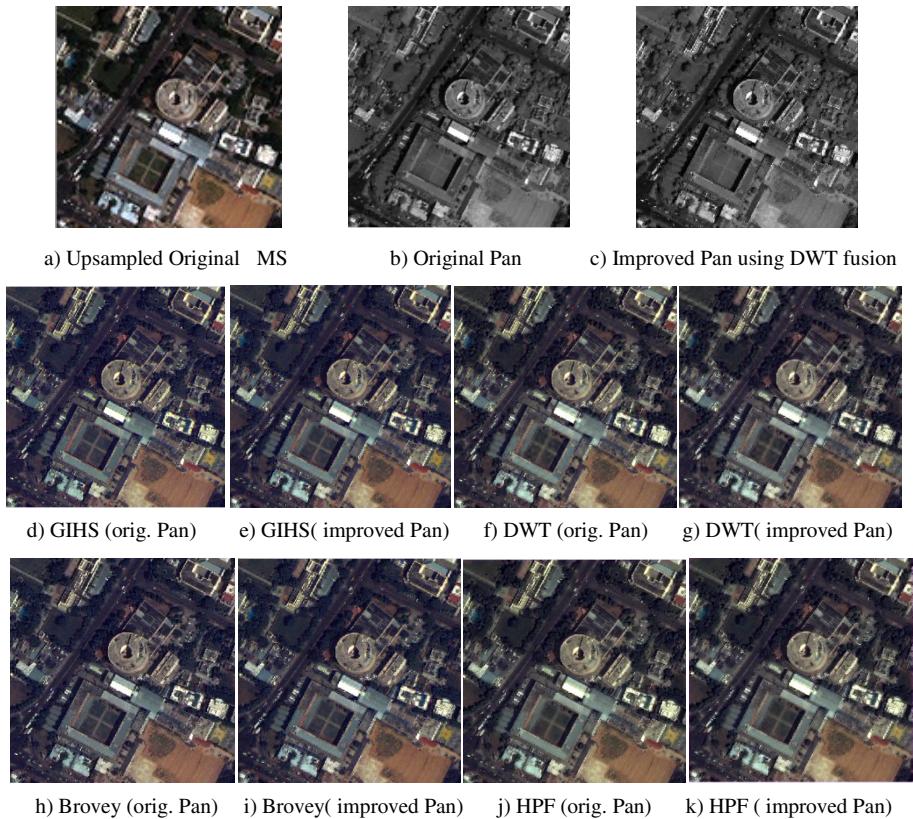
A high correlation between the merged filtered image and the Pan filtered one indicates that most spatial information of the Pan image has been incorporated during the fusion process.

## 5 Simulation and Results

The proposed method has been assessed on two very high-resolution image data sets collected by Quickbird and IKONOS satellites. Respectively, the first data set (original MS and original Pan shown in figure.4) has been acquired on the area of Jaipur, Rajasthan, India (Resolution: 2.4-m MS and 60-cm Pan. MS bands: R, G, B and NIR), the second data set (original MS and original Pan shown in figure.5) has been acquired on the area of São Paulo, Brazil (Resolution: 3.2 -m MS and 80-cm Pan. MS bands: R,G,B and NIR).

The comparisons are performed on both Pan-sharpening with original Pan and improved Pan images. We use here two fusion methods for Pan image improvement (in the frequency domain). The first set of images (Figure.4) shows different Pan-sharpening techniques using improved Pan based on DWT wavelet fusion. In the second set (figure.5) the improved Pan is based on High Pass Filtering (HPF) fusion. The pan-sharpening techniques used for assessment are: Generalized IHS, DWT Wavelet fusion, Brovey and High Pass Filtering (HPF). The comparison is performed in each technique using original Pan and improved Pan.

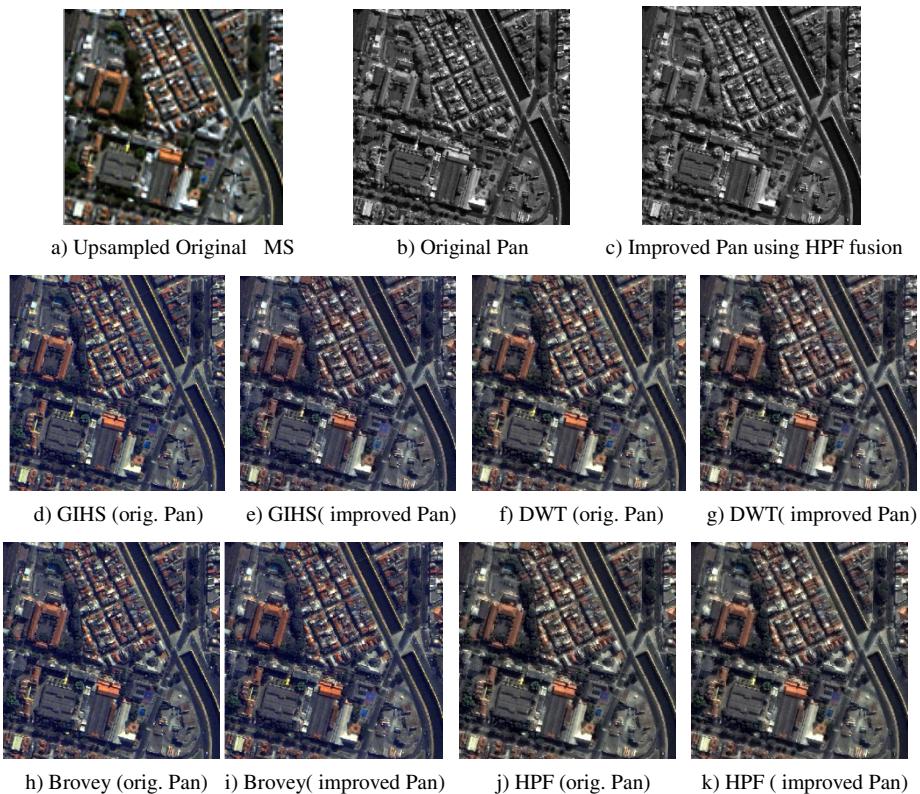
We can see that the proposed method gives better results than using original Pan in terms of spectral distortion and that preserves the spatial details well. The results in the frequency domain of pan-sharpening techniques such as DWT or HPF are slightly better on spectral distortion assessment using improved Pan compared with the original one. It is obvious that these techniques get better spectral quality. We can also see that the spatial quality is enhanced in this case using the proposed method.



**Fig. 4.** Pan-sharpening methods using original Pan and improved Pan based on DWT fusion

**Table 1.** Comparison of quality assessment parameters for Pan-sharpening techniques using original Pan and Improved Pan based on DWT fusion

Fusion methods	C C		ERGAS		RASE		RMSE		Spatial C	
	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan
GIHS	0.952	<b>0.975</b>	2.315	<b>1.931</b>	9.009	<b>7.552</b>	38.657	<b>32.408</b>	0.923	<b>0.925</b>
DWT	0.958	<b>0.967</b>	2.083	<b>1.881</b>	8.565	<b>7.803</b>	36.752	<b>33.484</b>	0.942	<b>0.949</b>
Brovey	0.954	<b>0.975</b>	2.293	<b>1.913</b>	9.014	<b>7.554</b>	38.682	<b>32.414</b>	0.924	<b>0.925</b>
HPF	0.961	<b>0.968</b>	2.077	<b>1.899</b>	8.556	<b>7.880</b>	36.715	<b>33.814</b>	0.977	<b>0.980</b>



**Fig. 5.** Pan-sharpening methods using original Pan and improved Pan based on HPF fusion

**Table 2.** Comparison of quality assessment parameters for Pan-sharpening techniques using original Pan and improved Pan based on HPF fusion

Fusion methods	CC		ERGAS		RASE		RMSE		Spatial C	
	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan	Pan	Improved Pan
GIHS	0.930	<b>0.955</b>	2.630	<b>2.324</b>	10.300	<b>8.089</b>	26.473	<b>23.360</b>	0.923	<b>0.930</b>
DWT	0.953	<b>0.958</b>	2.345	<b>2.243</b>	8.999	<b>8.588</b>	23.130	<b>22.074</b>	0.943	<b>0.951</b>
Brovey	0.930	<b>0.955</b>	2.626	<b>2.320</b>	10.299	<b>9.087</b>	26.471	<b>23.356</b>	0.932	<b>0.929</b>
HPF	0.959	<b>0.963</b>	2.329	<b>2.245</b>	8.912	<b>8.566</b>	22.907	<b>22.016</b>	0.976	<b>0.980</b>

## 6 Conclusion

Pan-sharpening of high resolution Pan image with low resolution MS image is an efficient way to get a high resolution MS image which is used in many fields of remote sensing. Unfortunately, the spectral quality is affected by most of Pan-sharpening techniques. Recently, many researchers attempt to deal with these issues. In our method, we improve the correlation between Pan and MS images by generating a new Pan image based on fusing the original Pan with the intensity image derived from MS image, in order to extract spectral details from MS and inject them into the Pan, and later, the pan-sharpening process will be performed on MS and the improved Pan to get better results than using original Pan.

**Acknowledgement.** This work is funded by the National Basic Research Program of China (No. 2010CB327902), the National Natural Science Foundation of China (No. 61375036, 61005016), the Beijing Natural Science Foundation (No. 4132064), the Program for New Century Excellent Talents in University, the Beijing Higher Education Young Elite Teacher Project, and the Fundamental Research Funds for the Central Universities. Zhaoxiang Zhang is the corresponding author of this paper.

## References

1. Amro, I., Mateos, J., Vega, M., Molina, R., Katsaggelos, A.K.: A survey of classical methods and new trends in pansharpening of multispectral images. *EURASIP Journal on Advances in Signal Processing* (2011)
2. Zhang, Y.: Understanding image fusion. *Photogrammetric Engineering and Remote Sensing* 70(6), 657–661 (2004)
3. Lee, S.H.: High-resolution reconstruction of multispectral imagery based on panchromatic imagery. In: *Proceedings of the IEEE Geosci. Remote Sens. Symp.*, vol. 2, pp. 101–104 (2008)
4. Shah, V.P., Younan, N.H., King, R.L.: An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets. *IEEE Trans. Geosci. Remote Sens.* 46, 1323–1335 (2008)
5. Palsson, F., Sveinsson, J.R., Ulfarsson, M.O.: A New Pansharpening Algorithm Based on Total Variation. *IEEE Geoscience and Remote Sensing Letters* 11(1), 318–322 (2014)
6. Liu, L., Wang, Y., Wang, Y.: Adaptive steepest descent method for Pan-sharpening of multispectral images. *Optical Engineering* 50(9) (September 2011)
7. Liu, L., Wang, Y., Wang, Y., Haiyan, Y.: Pansharpening using an adaptive linear model. In: *20th International Conference on Pattern Recognition*, pp. 4512–4515 (2010)
8. Guest editorial: Image fusion: Advances in the state of the art. *Information Fusion* 8, 114–118 (2007) ISSN 1566-2535
9. Wald, L., Ranchin, T., Mangolini, M.: Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering & Remote Sensing* 63(6), 691–699 (1997)
10. Tsagiris, V., Anastassopoulos, V.: An information measure for assessing pixel-level fusion methods. In: *Proc. SPIE*, vol. 5573, pp. 64–71 (2004)

11. Ranching, T., Wald, L.: Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. *Photogramm. Eng. Remote Sens.* 66, 49–61 (2000)
12. Wald, L.: Quality of high resolution synthesized images: Is there a simple criterion? In: *Proc. Int. Conf. Fusion Earth Data*, pp. 99–105 (January 2000)
13. Zhou, J., Civco, D.L., Silander, J.A.: A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* 19(4), 743–757 (1998)

# Combining SIFT and Individual Entropy Correlation Coefficient for Image Registration

Gan Liu<sup>1</sup>, Shengyong Chen<sup>1</sup>, Xiaolong Zhou<sup>1</sup>, Xiaoyan Wang<sup>1</sup>, Qiu Guan<sup>1</sup>,  
and Hui Yu<sup>2</sup>

<sup>1</sup>College of Computer Science and Technology  
Zhejiang University of Technology, Hangzhou, Zhejiang Province, China  
gan.lau@outlook.com, sy@ieee.org  
{zxl, xiaoyanwang, gq}@zjut.edu.cn  
<sup>2</sup> School of Computing, University of Portsmouth, Portsmouth, England  
hui.yu@port.ac.uk

**Abstract.** Image registration is an important topic in many fields including industrial image analysis systems, medical and remote sensing. To improve the registration accuracy, an image registration method that combines scale invariant feature transform and individual entropy correlation coefficient (SIFT-IECC) is proposed in this paper. First, scale invariant feature transform algorithm is applied to extract feature points to construct a transformation model. Then, a rough registration image is obtained according to the transformation model. The individual entropy correlation coefficient is used as the similarity measure to refine the rough registration image. Finally, the experimental results show the superior performance of the proposed SIFT-IECC registration method by comparing with the state-of-the-art methods.

**Keywords:** Image registration, Scale invariant feature transform, Individual entropy correlation coefficient.

## 1 Introduction

Image registration is the process of spatially aligning two or more images of the same scene acquired with, for example, different sensors or the same sensors at different times [1, 2, 3]. The registration geometrically aligns two images called the reference and floating images, respectively. Image registration has important applications in many fields including remote sensing [4], medical [5], and industrial image analysis systems [6]. In the field of computer vision, image registration is a critical component of image processing, such as image mosaicking [7], image fusion [8], image reconstruction [9] and so forth. Usually, image registration methods are generally categorized into two classes [1, 2]: feature-based method [3] and intensity-based method [6]. In general, the feature-based method is preferably applied when images contain many salient and detectable features, while the intensity-based method is recommended when images contain not enough features or the features are similar. However, in feature-based method, the process of features extracting is sensitive to noise, which

can easily lead the result that features in both reference image and floating image are hard to be detected and/or unstable in time. Therefore, it is hard to make a correspondence between the two feature sets in feature-based method. In intensity-based method, entire images have to be used during the registration steps which cause costly consumption of time and memory. Furthermore, nonlinear illumination changes exert negative effects on the registration result and therefore the intensity-based method does not perform well in stability which may lead to local extremum.

To improve the accuracy and stability, scale invariant feature transform [10] and individual entropy correlation coefficient [11] (SIFT-IECC) are combined to register the reference image and the floating image in this paper. Scale invariant feature transform (SIFT) is an algorithm to detect local features in images. The algorithm was first reported by David G. Lowe in 1999, it became a consummate algorithm till 2004. The SIFT algorithm has been applied in many fields like gesture recognition, object recognition, image stitching, video tracking, 3D modeling, matching moving and so on. In this paper, the SIFT algorithm is applied to extract the feature points that are used to construct a transformation model. A rough registration image is then obtained according to this transformation model. The SIFT-based affine transformation can correct the translation, rotation and scale of the floating image, so we use it first .To get the refined image registration image, the individual entropy correlation coefficient (IECC) is used as the similarity measure.

The flow chart of the proposed method is shown as in Fig. 1. The heavy line represents rough registration process, and thin line represents refining registration process. The corresponding steps of the proposed method are as follows. 1) Extract features of the reference and floating images using SIFT. 2) Match SIFT features of the reference and floating images. 3) Establish the affine transformation model based on the minimum mean square error (MMSE) method of matching feature points. 4) Obtain the rough registered image by transforming the floating image based on the affine transformation model. 5) Initialize the parameters of the IECC-based registration image. 6) Establish an affine transformation model by optimizing initializations or parameters, and then obtain the rough registered image by the affine transformation model. 7) Use IECC as the similarity measure to refine the rough registered image. 8) Output the refined registered image.

## 2 SIFT-IECC Image Registration

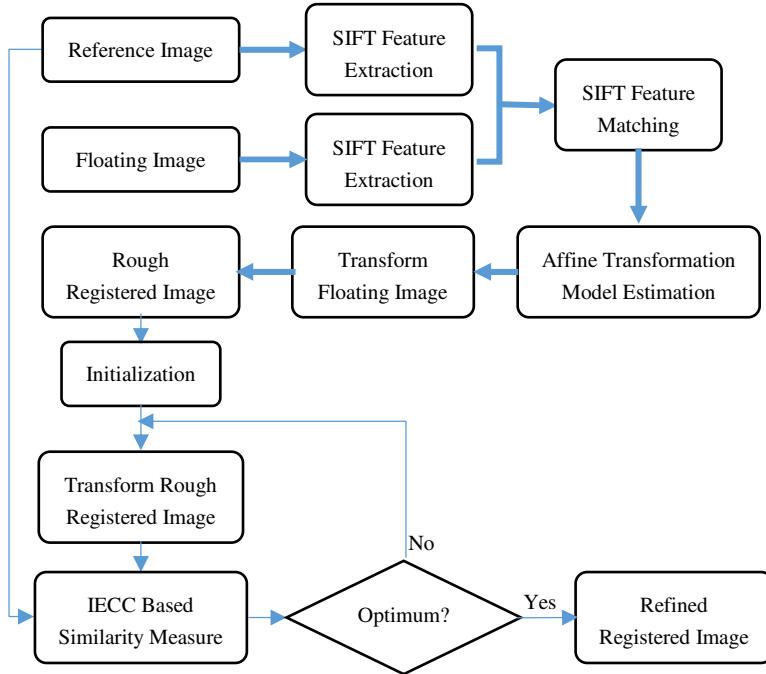
### 2.1 Feature Extraction Using SIFT

The major stages of the SIFT algorithm are stated in the following.

#### 1. Scale-space extrema detection

The Gaussian kernel has been proved to be the only possible kernel that can produce scale-space [12]. Therefore, the scale-space of an image is defined as a function produced from the convolution of a variable-scale Gaussian kernel with an input image.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$



**Fig. 1.** Flow chart of the proposed method

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

where  $L(x, y, \sigma)$  is the scale-space,  $G(x, y, \sigma)$  is the Gaussian kernel,  $I(x, y)$  is the input image,  $\sigma$  is the scale-space factor, and  $*$  is the convolution operation in  $x$  and  $y$ .

Scale-space extreme is used in the difference-of-Gaussian function convolved with the image for the purpose of detecting stable keypoint locations in scale-space efficiently.

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3)$$

where  $D(x, y, \sigma)$  is the difference-of-Gaussian function,  $k$  is a constant multiplicative factor.

Maxima and minima of the difference-of-Gaussian images are detected by comparing each sample point to its eight neighbors in the current image and nine neighbors in the scale above and below.

## 2. Keypoint localization

Points those are sensitive to noise or poorly localized along an edge should be rejected to pinpoint the local extreme, enhance the matching stability and improve noise

immunity.  $D(x,y,\sigma)$  is approximated by the Taylor expansion and it can be used to remove the unstable extrema with low contrast by discarding points whose offset is greater than an appropriate threshold. In order to eliminate edge responses, the ratio between the square of Hessian matrix's trace and Hessian matrix's determinant is used. An extremum with a large principal curvature across the edge but a small one in the perpendicular direction will be discarded.

### 3. Orientation assignment

Using local image properties to assign a consistent orientation to each extremum can guarantee invariance to image rotation. For each Gaussian smoothed image  $L(x,y)$ , the gradient magnitude  $m(x,y)$  and the orientation  $\theta(x,y)$  are pre-computed using pixel difference.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \quad (4)$$

$$\theta(x,y) = \tan^{-1} \left( \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \right) \quad (5)$$

### 4. Keypoint descriptor

Descriptors of the keypoints in both reference image and floating image are needed for matching. A keypoint descriptor is established by first calculating the gradient magnitude and orientation at every point in a region around the keypoint location, which means that the descriptor contains not only the keypoint's information but also pixels' information around the keypoint. These pixels around the keypoint are accumulated into orientation histograms to summarize the contents over 4x4 subregions that can form a descriptor of 8 orientations. 16x16 pixels around a keypoint is used to form 4x4 descriptors of 128 dimensional SIFT feature vector that is used to describe the keypoint, which will achieve invariance to image rotation optimally. To reduce the effects of illumination change, the SIFT feature vector is further normalized.

## 2.2 Feature Matching and Transformation Model Estimation

### 1. Feature Matching

Feature matching is used to establish an affine transformation model relying on the correspondences between features in both reference and floating image. An effective method to match feature points is to compare the distance of the nearest neighbor to that of the second nearest neighbor. More specially, with regard to each keypoint in the reference image, we can find the nearest keypoint with the shortest Euclidean distance,  $d_1$ , and the second nearest keypoint with second shortest Euclidean distance,  $d_2$ , in the floating image. If the ratio,  $d_1/d_2$ , is bigger than an appropriate threshold that can be confirmed by testing, the correspondence will be regarded as an incorrect match. K-d tree is used to improve the matching efficiency.

## 2. Transformation Model Estimation

After feature matching, we can get  $N$  matching point-pairs

$$\left\{ (x_{r,i}, y_{r,i}), (x_{f,i}, y_{f,i}) \right\}_{i=1,2,3,\dots,N} \quad (6)$$

where  $(x_{r,i}, y_{r,i})$  is the keypoint in the reference image and  $(x_{f,i}, y_{f,i})$  is the keypoint in the floating image.

An affine transformation model can be defined as:

$$(x_r, y_r, 1) = (x_f, y_f, 1) \cdot \begin{pmatrix} s \cdot \cos \alpha & s \cdot \sin \alpha & 0 \\ -s \cdot \sin \alpha & s \cdot \cos \alpha & 0 \\ t_x & t_y & 1 \end{pmatrix} \quad (7)$$

where  $s$  is the scale factor,  $\alpha$  is the rotation angle,  $t_x$  is the translation in the x-axis and  $t_y$  is the translation in the y-axis.

In this paper, MMSE is used to calculate the parameters of the affine transformation model.

$$s = \frac{\sum_i^N \sum_j^N \sqrt{(x_{r,i} - x_{r,j})^2 + (y_{r,i} - y_{r,j})^2}}{\sum_i^N \sum_j^N \sqrt{(x_{f,i} - x_{f,j})^2 + (y_{f,i} - y_{f,j})^2}} \quad (8)$$

$$\alpha = \frac{1}{N-1} \sum_{i=1}^N \sum_{j=1}^N \left( \tan^{-1} \frac{y_{r,i} - y_{r,j}}{x_{r,i} - x_{r,j}} - \tan^{-1} \frac{y_{f,i} - y_{f,j}}{x_{f,i} - x_{f,j}} \right) \quad (9)$$

$$t_x = \frac{1}{N} \sum_{i=1}^N (x_{r,i} - s \cdot \cos \alpha \cdot x_{f,i} + s \cdot \sin \alpha \cdot y_{f,i}) \quad (10)$$

$$t_y = \frac{1}{N} \sum_{i=1}^N (y_{r,i} - s \cdot \sin \alpha \cdot x_{f,i} - s \cdot \cos \alpha \cdot y_{f,i}) \quad (11)$$

Once the parameters confirmed, a rough registered image is obtained by transforming the floating image using affine transformation with bicubic interpolation. Now the rough registered image is regarded as the floating image so the next step is to register the reference image and the rough registered image.

### 2.3 IECC-Based Similarity Measure

In this paper, the similarity measure is IECC, which is a new similarity measure based on entropy. For two images R and F, we can calculate the marginal probability

distribution,  $p(r_i)$  and  $p(f_j)$ , the joint probability distribution,  $p(r_i, f_j)$ , of image R and F. The joint probability distribution  $p(r_i, f_j)$  can be obtained simply by normalizing the 2D histogram.

$$p(r_i, f_j) = \frac{h(r_i, f_j)}{\sum_{i=1}^{bin} \sum_{j=1}^{bin} h(r_i, f_j)} \quad (12)$$

where  $r_i$  is the intensity of image R,  $f_j$  is the intensity of image F,  $h(r_i, f_j)$  is the 2D histogram calculated from the two images R and F,  $bin$  is the size of the 2D histogram. The marginal probability  $p(r_i)$  and  $p(f_j)$  can be obtained by summing  $p(r_i, f_j)$  over  $f$  and  $r$ , respectively.

$$p(r_i) = \sum_{j=1}^{bin} p(r_i, f_j) \quad (13)$$

$$p(f_j) = \sum_{i=1}^{bin} p(r_i, f_j) \quad (14)$$

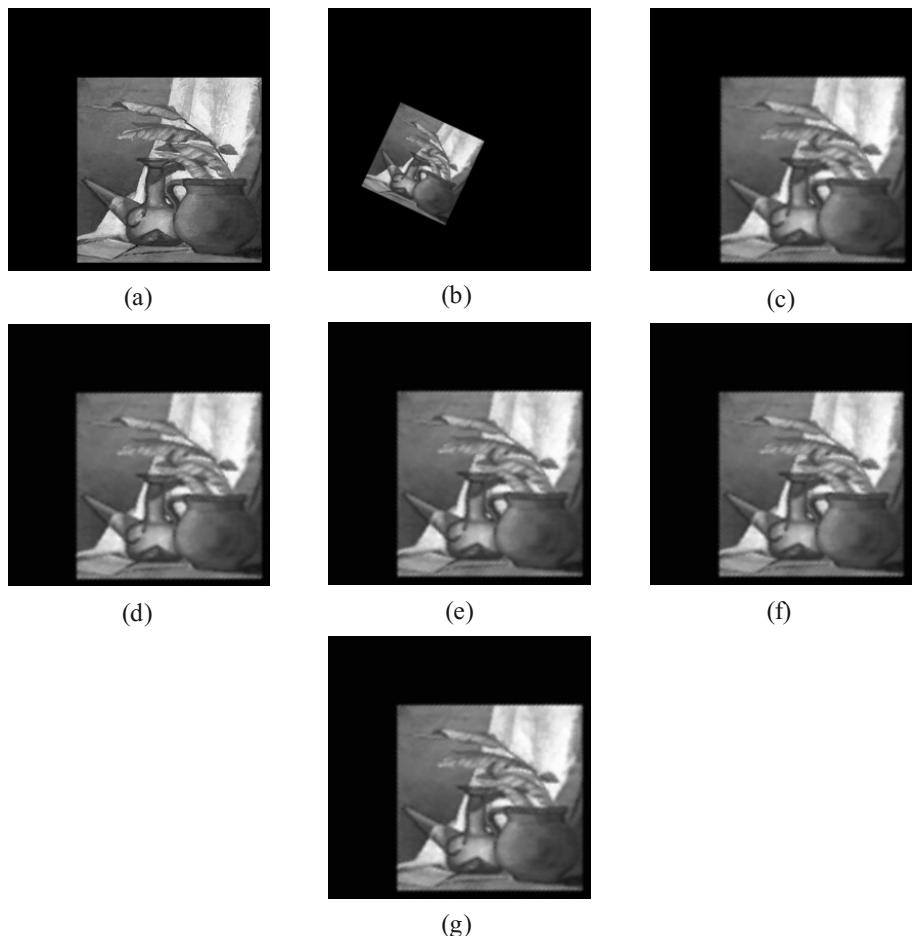
IECC can be defined as:

$$\text{IECC}(R, F) = - \sum_{i=1}^{bin} \sum_{j=1}^{bin} \frac{p(r_i, f_j) \cdot \log_2 \left( p(r_i, f_j) / (p(r_i) \cdot p(f_j)) \right)}{p(r_i) \cdot \log_2(p(r_i)) + p(f_j) \cdot \log_2(p(f_j))} \quad (15)$$

IECC is used to determine if the reference image and the floating image are registered. The maximum IECC is determined as the best registration between two images.

### 3 Experiments

In this section, experiments are conducted to evaluate the performance of the proposed image registration method. There are many similarity measures such as mutual information (MI) [5], normalized mutual information (NMI) [5] and entropy correlation coefficient (ECC) [13] that can be combined with SIFT to register the image. The proposed SIFT-IECC image registration method is compared with the SIFT [14] method, the SIFT-MI method [15], the SIFT-NMI method and the SIFT-ECC method. The experiments are implemented in Matlab 2012a using a computer with a CPU of Inter Core i5 (2.5GHz) and 4GB memory.



**Fig. 2.** The image registration results of image 1. (a) Reference image. (b) Floating image. (c) SIFT registration result. (d) SIFT-MI registration result. (e) SIFT-NMI registration result. (f) SIFT-ECC registration result. (g) SIFT-IECC registration result.

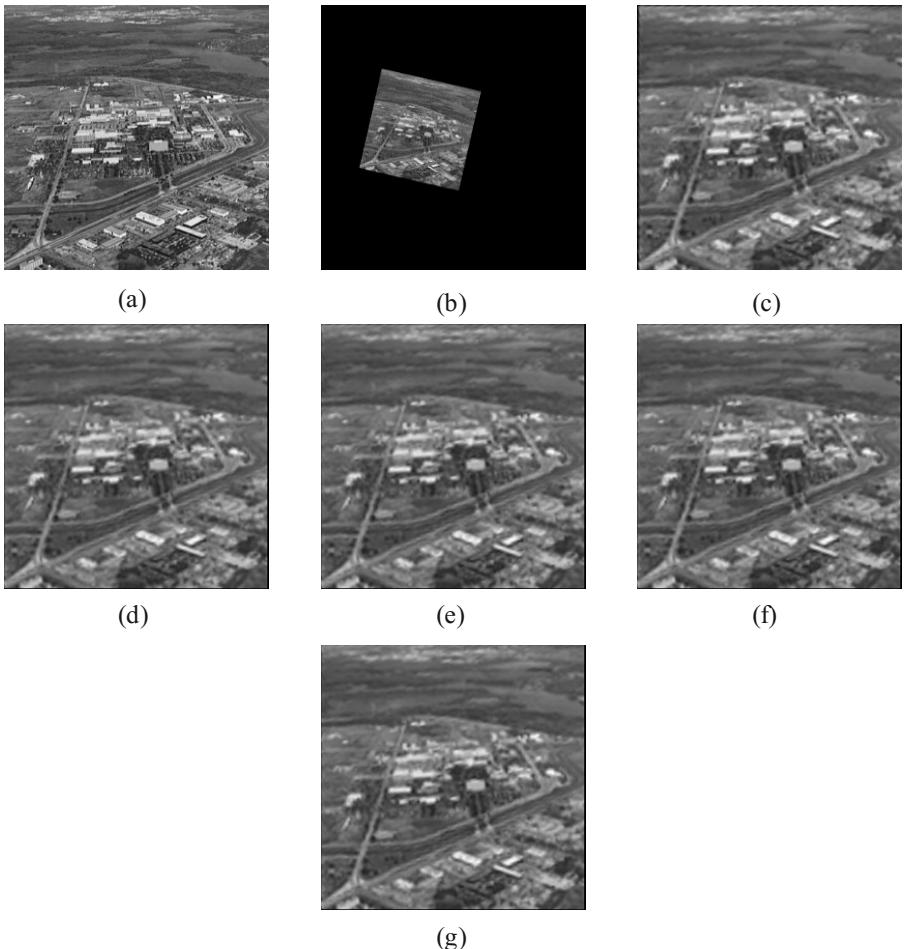
**Table 1.** The registration results of image 1

Type of Method	$t_x$ (pixel)	$t_y$ (pixel)	$\alpha$ (rad)	$s$	RMSE (pixel)	Time (second)
SIFT	-137.2329	-39.3152	-0.4356	2.0006	1.6936	85.6430
SIFT-MI	-139.8706	-40.0927	-0.4327	2.0109	0.4917	112.3408
SIFT-NMI	-138.8549	-39.3212	-0.4340	2.0010	0.3272	174.1210
SIFT-ECC	-138.8549	-39.3212	-0.4340	2.0010	0.3272	180.9565
SIFT-IECC	-139.0405	-38.4798	-0.4372	2.0031	0.3192	162.1692

The registration efficiency is assessed by the computation time, and the registration accuracy is assessed by the root mean square error (RMSE) of the detected key points. The smaller RMSE indicates the better registration result. The RMSE between the reference image and the registered image can be defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( (R_{x_i} - Reg_{x_i})^2 + (R_{y_i} - Reg_{y_i})^2 \right)} \quad (16)$$

where  $N$  is the number of key points of the reference image and the registered image,  $R_{x_i}$  and  $R_{y_i}$  are key points coordinates of the reference image,  $Reg_{x_i}$  and  $Reg_{y_i}$  are key points coordinates of the registered image.



**Fig. 3.** The image registration results of image 2. (a) Reference image. (b) Floating image. (c) SIFT registration result. (d) SIFT-MI registration result. (e) SIFT-NMI registration result. (f) SIFT-ECC registration result. (g) SIFT-IECC registration result.

**Table 2.** The registration results of image 2

Type of Method	$t_x$ (pixel)	$t_y$ (pixel)	$\alpha$ (rad)	s	RMSE (pixel)	Time (second)
SIFT	-184.1715	-122.3788	-0.2228	2.1743	1.8006	54.2974
SIFT-MI	-185.9644	-122.1512	-0.2228	2.1751	0.3006	223.8533
SIFT-NMI	-185.9236	-122.1541	-0.2229	2.1752	0.2951	183.2656
SIFT-ECC	-185.9236	-122.1541	-0.2229	2.1752	0.2951	182.6300
SIFT-IECC	-185.9054	-122.0525	-0.2229	2.1741	0.2940	183.8825

Two images obtained from [www.prenhall.com/gonzalezwoods](http://www.prenhall.com/gonzalezwoods) are used to test the above methods. The reference and floating images are re-sampled to 256x256 for computation convenience. The registration results are shown in Fig. 2, Fig. 3, Tab. 1 and Tab. 2, respectively. In the table,  $t_x$  and  $t_y$  are the translations in x-axis and y-axis respectively,  $\alpha$  is the rotation angle and the s is the scale.

The results demonstrate that the proposed SIFT-IECC image registration method provides a significant improvement in RMSE. Moreover, the processing time of the SIFT-IECC image registration method is comparable to the SIFT-MI method, the SIFT-NMI method and the SIFT-ECC method. It costs more time than the SIFT registration method, because it incorporates the IECC-based similarity measure.

## 4 Conclusion

In this paper, the SIFT-IECC registration method is proposed. The SIFT algorithm is used to obtain an affine transformation model and get a rough registered image. The reference image and the rough registered image are refined using IECC. The registration results demonstrate that the proposed SIFT-IECC registration method performs better than the SIFT method, the SIFT-MI method, the SIFT-NMI method and the SIFT-ECC method in terms of accuracy. Our future work will focus on improving the efficiency.

**Acknowledgements.** This work was partially supported by the National Science Foundation of China (NSFC NO. 61273286, 61325019, 61173096, 11302195 and 61103140).

## References

1. Xing, C., Qiu, P.H.: Intensity-Based Image Registration by Nonparametric Local Smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(10), 2081–2092 (2011)
2. Oliveira, F.P.M., Tavares, J.: Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering* 17(2), 73–93 (2014)

3. Cheng, D., Xie, S.Q., Hammerle, E.: A Robust Local Descriptor Method for Registering Maori Artefacts using Colour Images. In: International Conference on Information and Automation, vols. 1-3. IEEE, New York (2009)
4. Liang, J., Liu, X., Huang, K., Li, X., Wang, D., Wang, X.: Automatic Registration of Multisensor Images Using an Integrated Spatial and Mutual Information (SMI) Metric. *IEEE Transactions on Geoscience and Remote Sensing* 52(1), 603–615 (2014)
5. Yokoi, T., Soma, T., Shinohara, H., Matsuda, H.: Accuracy and reproducibility of co-registration techniques based on mutual information and normalized mutual information for MRI and SPECT brain images. *Annals of Nuclear Medicine* 18(8), 659–667 (2004)
6. Xu, H.L., Hua, G.R., Zhuang, J., Wang, S.A.: A Frequency Domain Approach to Fast and Accurate Image Registration. In: International Conference on Information and Automation, vols. 1-3. IEEE, New York (2009)
7. Hurtos, N., Cuf, X., Petillot, Y., Salvi, J., Robotics Society of, J.: Fourier-based Registrations for Two-Dimensional Forward-Looking Sonar Image Mosaicing. In: 25th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5298–5305. IEEE (2012)
8. Goshtasby, A.A.: 2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications. Wiley (2005)
9. Lee, E.S., Kang, M.G.: Regularized adaptive high-resolution image reconstruction considering inaccurate subpixel registration. *IEEE Transactions on Image Processing* 12(7), 826–837 (2003)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
11. Itou, T., Shinohara, H., Sakaguchi, K., Hashimoto, T., Yokoi, T., Souma, T.: Multimodal image registration using IECC as the similarity measure. *Medical Physics* 38(2), 1103–1115 (2011)
12. Lindeberg, T.: Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics* 21(1-2), 225–270 (1994)
13. Skerl, D., Likar, B., Fitzpatrick, J.M., Pernus, F.: Comparative evaluation of similarity measures for the rigid registration of multi-modal head images. *Physics in Medicine and Biology* 52(18), 5587–5601 (2007)
14. Moradi, M., Abolmaesumi, P.: Medical image registration based on distinctive image features from scale-invariant (SIFT) key-points. In: 19th International Congress and Exhibition on Computer Assisted Radiology and Surgery, vol. 1281, pp. 1292–1292. Elsevier (2005)
15. Suri, S., Schwind, P., Reinartz, P., Uhl, J.: Combining mutual information and scale invariant feature transform for fast and robust multisensor SAR image registration. In: Proceedings of the 75 ASPRS Annual Conference (2009)

# Spectral Fidelity Analysis of Compressed Sensing Reconstruction Hyperspectral Remote Sensing Image Based on Wavelet Transformation<sup>\*</sup>

Yi Ma<sup>\*\*</sup>, Jie Zhang, and Ni An

First Institute of Oceanography, State Oceanic Administration, Qingdao, China  
mayimail@fio.org.cn

**Abstract.** For hyperspectral image research, spectral characteristic retainment is more important than the spatial details retainment, so it is necessary to evaluate the spectral influence of hyperspectral image compressed sensing. In this paper, the researchers select a hyperspectral remote sensing image PROBE CHRIS with abundant coastal wetland ground objects to analyze spectral fidelity of wavelet transform compressed sensing algorithm on the basis of three indicators between reconstruction and original image pixel spectra: correlation coefficient, error and relative error. Meanwhile, eight typical ground objects are chosen to analyze their respective spectral deviation. The results indicate: (1) Image reconstruction algorithm based on wavelet transform compressed sensing functions well. Between the pixels of reconstruction image and original one, their average spectral correlation coefficient is 0.9428, error is 6.4096, and relative error is 13.81%; (2) Spectrum fidelity indicator values vary with wavebands. Reconstruction algorithm is selective about objects.

**Keywords:** Spectral fidelity, compressed sensing reconstruction, hyperspectral remote sensing image.

## 1 Introduction

Hyperspectral remote sensing data include both image information and spectral information. It can obtain continuous spectral curves of each pixel while imaging, so it is suitable for quantitative remote sensing, fine classification and target detection. However Along with the increasing abundance of spectral information from hyperspectral remote sensing data, the data quantity increases dramatically. This demands more effective data storage and transmission. Compressed sensing (CS) measures and encodes the image projection value from higher dimension to lower dimension with the sampling rate far below Nyquist. Its decoding process is not the conventional reverse process, but the accurate or approximate reconstruction of images based on

---

<sup>\*</sup> This paper is funded by National Science Fund of China (ID: 41206172) and Dragon Project III (ID: 10470)

<sup>\*\*</sup> Corresponding author.

signal sparse decomposition theory with blind source separation reverse. Compressed sensing is an image reconstruction algorithm which features low storage and transmission data quantity and excellent image restoration, but the pixel spectrum of the restored images will change. However, for hyperspectral image research, spectral characteristic retainment is more important than the spatial details retainment, so it is necessary to evaluate the spectral influence of hyperspectral image compressed sensing.

Compressed Sensing (CS) theory is put forward in 2006 by D. Donoho[1], E. Candes and J. Romberg[1, 2, 3]. From then on the CS theory has undergone rapid development and been applied in photography [4], medicine [5], face recognition [6], geophysics [7] and remote sensing [8]. It is a valuable vehicle of remote sensing image compression, transmission and reconstruction. Spectral fidelity research has focused on image fusion in recent years [9, 10, 11, 12], especially that of high spatial resolution and multispectral image data. In the research, the emphasis is on the spatial information integration and spectral information fidelity, but spectral fidelity of CS hyperspectral image is not discussed in detail. In fact, the spatial-spectral resolution increase and the image coverage expansion result in the exponential increase of hyperspectral remote sensing data quantity, so it is necessary to study the remote sensing image compression, transmission and reconstruction method with high spectral fidelity.

In this paper, the researchers select hyperspectral remote sensing image PROBE CHRIS with abundant coastal wetland ground objects to analyze spectral fidelity of wavelet transform compressed sensing algorithm on the basis of three indicators between reconstruction and original image pixel spectra: correlation coefficient, error and relative error. Meanwhile, eight typical ground objects are chosen to analyze their respective spectral deviation. The eight ground objects include Phragmites australis, Suaeda glauca, Taramix chinensis, Spartina, Salix babylonica, tidal flat, river and aquaculture water.

## 2 Hyperspectral Remote Sensing Data and Data Processing

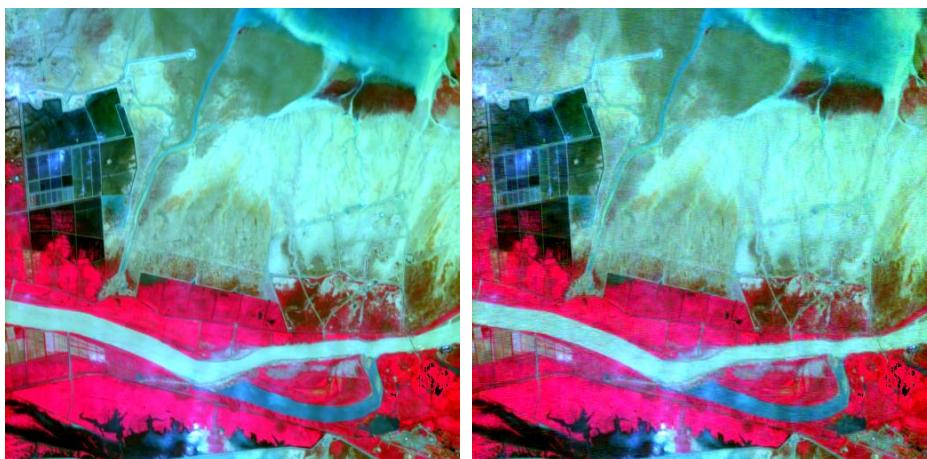
### 2.1 Hyperspectral Remote Sensing Data

CHRIS is a remote sensor loaded on ESA PROBA. Its full name is Compact High Resolution Imaging Spectrometer. CHRIS have five imaging modes(Table 1), and can gather remote sensing images from five angles, i.e.  $0^\circ$ ,  $+36^\circ$ ,  $-36^\circ$ ,  $+55^\circ$  and  $-55^\circ$ . This paper uses a Yellow river estuary CHRIS mode 2 image data (Figure 1) obtained in 2012 June, whose spectral range is 406-1035nm, spectral resolution 1.25-11.00nm and spatial resolution 17m.

The area where the CHRIS image covers is located in the junction between new and old Yellow river estuaries, which feature natural and artificial wetlands, such as Phragmites australis, Suaeda glauca, Taramix chinensis, Spartina, Salix babylonica, tidal flat, river and aquaculture water. Details of the typical ground objects are shown in Table 2.

**Table 1.** Specification of CHRIS

Mode	Spectral range(nm)	Band number	Spatial resolution(m)	Coverage(km <sup>2</sup> )	Application
1	406-1003	62	34	14×14	Land and water
2	406-1036	18	17	14×14	Water
3	438-1035	18	17	14×14	Land
4	486-796	18	17	14×14	Vegetation
5	438-1036	37	17	14×7	Land

**Fig. 1.** CHRIS hyperspectral image (left) and compressed sensing reconstruction image (right)

## 2.2 Data Processing

CHRIS is a push broom imaging sensor, whose striped noise originates from difference in spectral response functions of the each CCD unit brought about by its respective optical properties. In the paper, the researchers use Software HDFclean provided by ESA to remove the noise and fill some missing pixel. As a result, the vertical and horizontal striping noises are both eliminated in the CHRIS hyperspectral image, and the image has clear texture and conspicuous boundary with few changes in gray value.

The signal received by the remote sensor includes not only the surface signal of objects, but also the atmospheric contribution. Solar radiation is reduced by scattering and absorption of atmospheric aerosol and molecule, meanwhile, the signal received by the remote sensor is enhanced by some of direct or indirect scattering signals. So it is necessary to remove the atmospheric contribution from the total remote sensing signal, and this preprocessing is atmosphere correction. Software BEAM provided by ESA is used to do this, and the data of the corrected image are more accurate.

**Table 2.** Details of typical ground objects in Yellow River estuary

Ground object name	Subimage	Explain
<i>Phragmites austrialis</i>		<i>Phragmites austrialis</i> grow in fresh water and saline water, but grow better in the fresh water.
<i>Suaeda glauca</i>		<i>Suaeda glauca</i> is salt tolerant vegetation with sparse distribution, and grow in inter tidal zone.
<i>Taramix chinensis</i>		<i>Taramix chinensis</i> is salt tolerant shrub with sparse distribution, and grow in sandy or silt tidal flat and the coast.
<i>Spartina</i>		<i>Spartina</i> is a pioneer species in coastal wetland, and grow in tidal flat where tide often arrives.
<i>Salix babylonica</i>		<i>Salix babylonica</i> belongs to macrophanerophytes, and grow along the river channel.
<i>Tidal flat</i>		<i>Tidal flat</i> is the tide invasion zone between high tide and low tide. In this paper it refers to the mud flat.
<i>River</i>		<i>River</i> refers to the river near Yellow river estuary.
<i>Aquaculture water</i>		<i>Aquaculture water</i> refers to the artificial excavating or natural form for aquaculture pond.

### 3 Method

#### 3.1 Wavelet Compressed Sensing

Shannon Nyquist sampling theorem points out that the sampling rate must be at least twice the maximum frequency present in the signal. However, the sampling method may bring about data redundancy when massive information and high signal frequency are involved; meanwhile, hardware cost is also a challenge. Compressed sensing theory holds that signal can be measured with a frequency far lower than that of the original signal by a non-full rank matrix independent of signal expression bases if the signal is K-sparse in a transformation domain. The signal is reconstructed through solving a convex optimization problem. On balance, compressed sensing has the advantages of sparse sampling and simple encoding.

Supposed that image  $\mathbf{X}$  can be expressed as the linear combination of wavelet base  $\Psi = [\psi_1, \psi_2, \dots, \psi_m]$ , that is

$$\mathbf{X} = \sum_{k=1}^n \psi_k \alpha_k = \Psi \boldsymbol{\alpha} \quad (1)$$

When the image  $\mathbf{X}$  has  $k \ll n$  nonzero coefficient  $\alpha_k$ ,  $\Psi$  is considered as the expression base of  $\mathbf{X}$ .

The image  $\mathbf{X}$  is projected on  $\mathbf{Y}$  by Hadamard measured matrix  $\Phi = [\phi_1, \phi_2, \dots, \phi_m]$ , i.e.

$$\mathbf{Y} = \Phi \mathbf{X} \quad (2)$$

Substitute equation (1) into equation (2), then

$$\mathbf{Y} = \Phi \Psi \boldsymbol{\alpha} \quad (3)$$

As  $\boldsymbol{\alpha}$  is K-sparse and  $k \ll n$ , it can be calculated according to equation (3) by sparse decomposition algorithm if expression base  $\Psi$  and measured matrix  $\Phi$  are linearly independent.

The processes of wavelet compressed sensing image reconstruction include calculation of projection  $\mathbf{Y}$  through  $\Phi$  and reconstruction of image  $\bar{\mathbf{X}}$  by taking  $\boldsymbol{\alpha}$  into equation (1). The key to image reconstruction is the solving of sparse coefficient  $\boldsymbol{\alpha}$ , and  $\boldsymbol{\alpha}$  is the solution of  $l_1$  minimum norm optimization problem

$$\min \|\boldsymbol{\alpha}\|_{l_1} \quad s.t. \quad \mathbf{Y} = \Phi \Psi \boldsymbol{\alpha} \quad (4)$$

The solving algorithm of the above optimization problem is orthogonal matching pursuit method (OMP).

### 3.2 Evaluation Method of Spectral Fidelity

Three evaluation indicators that is spectral information retainment, spectral information distortion and spectral information matching are used to analyze spectral fidelity of compressed sensing reconstruction image. Supposed that  $F(i, j, k)$  and  $G(i, j, k)$  are the pixel gray values of original image  $\mathbf{X}$  and reconstruction image  $\bar{\mathbf{X}}$  respectively, where  $k \in \{1, 2, \dots, K\}$  is band number,  $i \in \{1, 2, \dots, M\}$  is row and  $j \in \{1, 2, \dots, N\}$  is column.

- Correlation coefficient is an average value of correlation coefficient between pixel spectra of reconstruction image and that of original image. It reveals the spectral

information retainment of reconstruction hyperspectral image data. The expression of correlation coefficient  $P$  is

$$P = \frac{\sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^L \frac{(F(i, j, k) - \bar{F}(i, j, k))(G(i, j, k) - \bar{G}(i, j, k))}{\sqrt{(F(i, j, k) - \bar{F}(i, j, k))^2(G(i, j, k) - \bar{G}(i, j, k))^2}}}{MN} \quad (5)$$

- Error is an average value of error between pixel spectra of reconstruction image and that of original image. It means the spectral information distortion of reconstruction hyperspectral image data. The expression of error  $D$  is

$$D = \frac{\sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^L |F(i, j, k) - G(i, j, k)|}{MNL} \quad (6)$$

- Relative error is an average value of relative error between pixel spectra of reconstruction image and that of original image. It discloses the spectral information mismatching degree between reconstruction image and original image. The expression of error  $D'$  is

$$D' = \frac{\sum_{i=1}^M \sum_{j=1}^N \sum_{k=1}^L \frac{|F(i, j, k) - G(i, j, k)|}{G(i, j, k)}}{MNL} \quad (7)$$

## 4 Results and Discussions

### 4.1 Reconstruction Image Based on Compressed Sensing

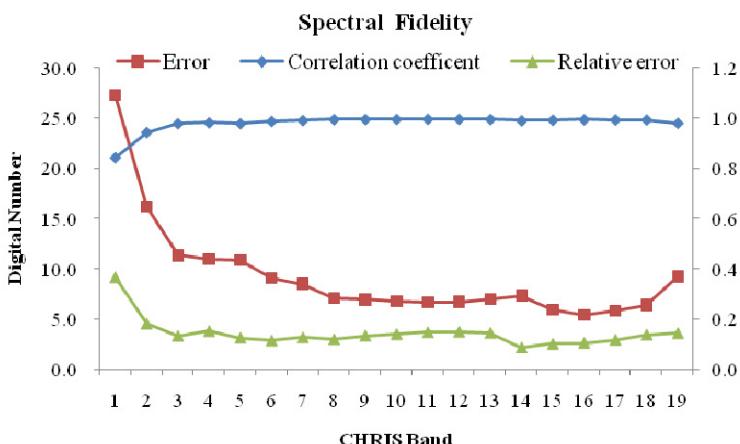
The CHRIS hyperspectral remote sensing image of the Yellow River estuary is reconstructed based on wavelet compressed sensing. When the measured matrix sample is taken as 300 after experiments, the reconstruction image is shown in Figure 1. Visually, the reconstruction image not only retain the information about large ground objects such as Phragmites australis and Tidal flat, but also the details of the small scattered objects such as Suaeda glauca and Taramix chinensis. According to the correlation coefficient and peak signal-to-noise ratio (PNSR) between the reconstruction image and original image of a single band (Table 3), most correlation coefficients among 18 bands are more than 0.95, and all the PNSR are more than 20. This shows that the image reconstructing has achieved the desirable outcome.

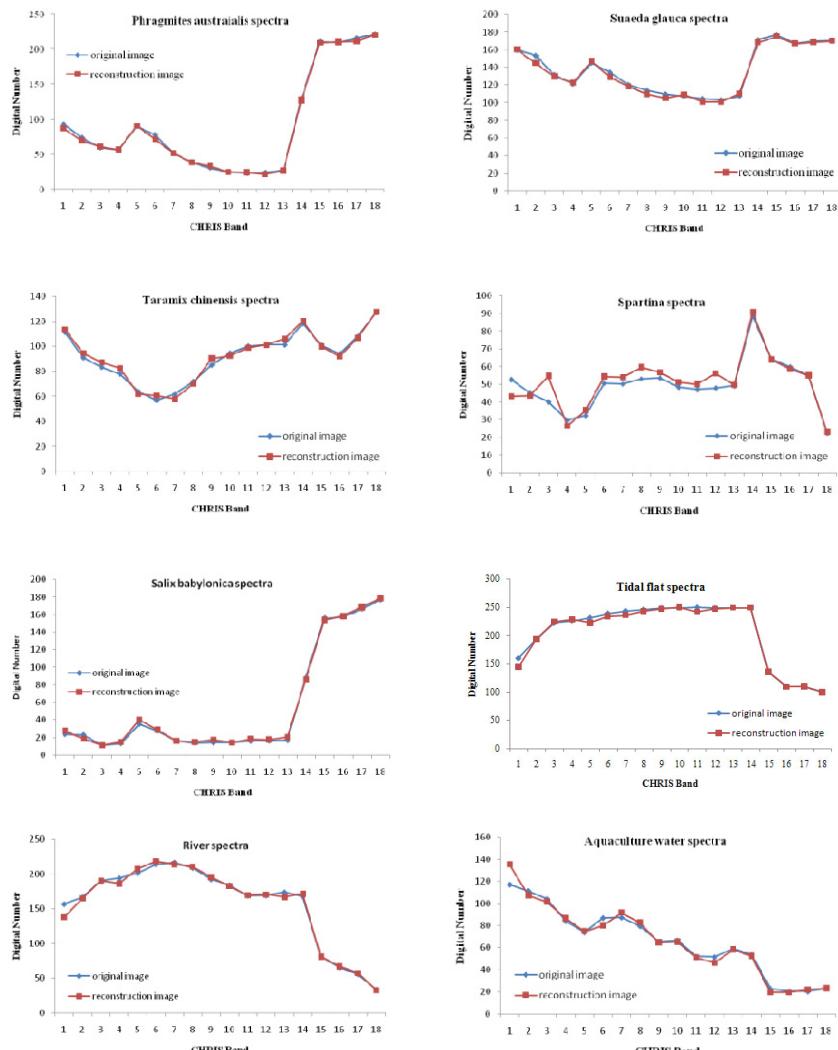
**Table 3.** Correlation coefficient and PNSR of reconstruction and original image

Correlation coefficient				PNSR			
1	0.8406	10	0.9939	1	17.1538	10	29.1040
2	0.9413	11	0.9942	2	21.766	11	29.2995
3	0.9757	12	0.9941	3	24.6758	12	29.1953
4	0.9803	13	0.9935	4	25.1458	13	28.8884
5	0.9765	14	0.9879	5	25.1642	14	28.1980
6	0.9855	15	0.9907	6	26.717	15	29.9224
7	0.9891	16	0.9924	7	27.2142	16	30.6364
8	0.9928	17	0.9916	8	28.8038	17	29.9382
9	0.9934	18	0.9915	9	28.9278	18	29.3560

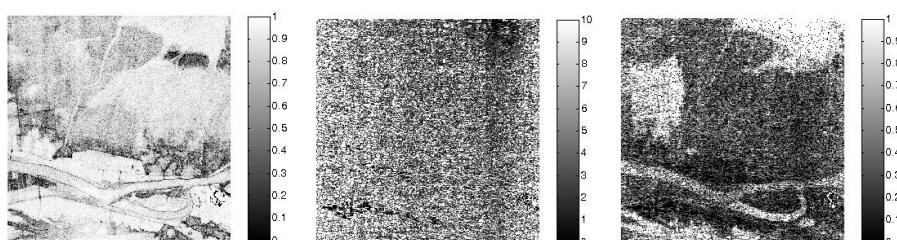
## 4.2 Analysis of Spectral Fidelity

Average correlation coefficient between the reconstruction and original image pixels is 0.9428, which exhibits high similarity between average spectra of two image data. This means the reconstruction image spectra retains its features well. The average error is 6.4096, indicating that the spectra have a small deviation between two image data, which shows that the reconstruction image spectra has less distortion. The relative error is 13.81%, denoting that the reconstruction image spectra have a small deviation proportion, which implies that the reconstruction image spectra match the original image spectra well.

**Fig. 2.** Three spectral fidelity index value in different band



**Fig. 3.** Three spectral fidelity index value in different band of 8 typical ground objects



**Fig. 4.** Three spectral fidelity index images

**Table 4.** Spectral fidelity index value of typical ground objects

Ground objects	Correlation coefficient	Error	Relative error
Phragmites austrialis	0.9834	5.3556	0.0251
Suaeda glauca	0.9283	5.8713	0.0541
Taramix chinensis	0.8676	7.0088	0.0579
Spartina	0.7878	3.5475	0.2772
Salix babylonica	0.9806	5.1200	0.0294
Tidal flat	0.9808	4.7450	0.0489
River	0.9833	6.4222	0.1433
Aquaculture water	0.9467	7.7213	0.4654

Spectrum fidelity indicator values are different in different wavebands. Three indicator values of 18 CHRIS wavebands are shown in Figure 2. The correlation coefficients are more than 0.8 in all bands; correlation coefficients of green, red and near infrared spectra are greater than 0.9; relative errors are smaller than 20% except that of the first band. To sum up, the spectral fidelity of green, red and near infrared band is better than that of the blue band.

The compressed sensing reconstruction image spectra change with ground objects. Figure 3 illustrates the spectral difference between original images and reconstruction images of 8 typical ground objects: Phragmites austrialis, Suaeda glauca, Taramix chinensis, Spartina, Salix babylonica, Tidal flat, river and aquaculture water. The comparison reveals that spectra fidelity is well persevered except for the Spartina spectra of the blue band. Figure 4 illustrates the spatial distribution images of three spectral fidelity indicators: correlation coefficient, error and relative error. The image pixels value indicates degree of spectral fidelity. Tab. 4 gives the spectral fidelity indexes value of the typical ground objects. According to Figure 4 and Table 4, the spectral fidelity of Phragmites austrialis is well kept, with the correlation coefficient being 0.9834; the spectral fidelity of River, Spartina and Salix babylonica come second, with the correlation coefficient being 0.9833, 0.9808 and 0.9806 respectively; the spectral fidelity of aquaculture water and Suaeda glauca come third with the correlation coefficient being 0.9467 and 0.9283; the spectral fidelity of Taramix chinensis and Spartina are the least ideal, with the correlation coefficient being 0.8676 and 0.7878. The error value of Spartina, 3.5475, is the smallest, indicating the slightest spectrum distortion; the error values of Tidal flat, Salix babylonica and Phragmites austrialis are respectively 4.7450, 5.12 and 5.3556; the error values of the Suaeda glauca, river and Taramix chinensis are respectively 5.8713, 6.4222 and 7.0088. The error value of aquaculture water is the highest, being 7.7213. This indicates the greatest spectral change. The relative error values of Phragmites austrialis and Salix babylonica are smaller, being 0.0251 and 0.0294. This means slight spectral changes.

Relative error values of Tidal flat, Suaeda glauca and Taramix chinensis are respectively 0.0489, 0.0541 and 0.0579. The relative errors of Salix babylonica and Aquaculture water are the highest, being 0.2772 and 0.4654. This means the low spectral fidelity. A comprehensive analysis of the three indicators reveals that among the chosen ground objects, Phragmites australis has the highest spectra fidelity, Spartina spectra have a largest deviation proportion, and Aquaculture water spectra have a lowest matching.

## 5 Conclusions

In this paper, the researchers select a hyperspectral remote sensing image PROBE CHRIS with abundant coastal wetland ground objects to analyze spectral fidelity of wavelet transform compressed sensing algorithm on the basis of three indicators between reconstruction and original image pixel spectra: correlation coefficient, error and relative error. Meanwhile, eight typical ground objects are chosen to analyze their respective spectral deviation. The results indicate: (1) in terms of spectral fidelity, image reconstruction algorithm based on wavelet transform compressed sensing functions well. Between the pixels of reconstruction image and original one, their average spectral correlation coefficient is 0.9428, error is 6.4096, and relative error is 13.81%; (2) spectrum fidelity indicator values vary with wavebands. The correlation coefficients are more than 0.8 in all bands, and those of green, red and near infrared spectral bands are more than 0.9. The relative errors are less than 20%. That is to say, the spectral fidelity of green, red and near infrared bands is better than that of the blue band. (3) Reconstruction algorithm is selective about objects. The three indicators reveal that Phragmites spectra is preserved best, Spartina spectra undergoes the greatest changes and Aquaculture water spectral matching is not as good as the others.

**Acknowledgements.** Thanks ESA for PROBE CHRIS data.

## References

1. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory* 52(2), 489–509 (2006)
2. Candès, E., Tao, T.: Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory* 52(12), 5406–5425 (2006)
3. Donoho, D.: Compressed sensing. *IEEE Transactions on Information Theory* 52(4), 1289–1306 (2006)
4. Duarte, M.F.: Single Pixel Imaging via Compressive Sampling (Building simpler, smaller, and less-expensive digital cameras). *IEEE Signal Processing Magazine* 25(2), 83–91 (2008)
5. Lustig, M.: Compressed sensing MRI. *IEEE Signal Processing Magazine* 25(2), 72–82 (2008)
6. Wright, J.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)

7. Hennenfent, G., Herrmann, F.J.: Simply denoise: wavefield reconstruction via jittered undersampling. *Geophysics* 73(3), 19–28 (2008)
8. Baraniuk, R., Steeghs, P.C.: Compressive radar imaging. In: 2007 IEEE Radar Conference, Boston, pp. 128–133 (2007)
9. Zhang, Y.: Understanding image fusion. *Photogrammetric Engineering and Remote Sensing* 70(6), 657–661 (2004)
10. Li, C., Xu, H.: Spectral Fidelity in High-resolution Remote Sensing Image Fusion. *Geo-information Science* 10(4), 520–526 (2008)
11. Wang, J., Wu, L.: An image fusion algorithm for spectrum respective based on wavelet. *Science of Surveying and Mapping* 35(5), 120–122 (2010)
12. Yang, K., Zhang, T., Wang, L., Qian, X., Wang, L., Liu, S.: Harmonic Analysis Fusion of Hyperspectral Image and Its Spectral Information Fidelity Evaluation. *Spectroscopy and Spectral Analysis* 33(9), 2496–2501 (2013)

# A Fast Algorithm for Image Defogging

Xiaoyan He<sup>1</sup>, Jianxu Mao<sup>1</sup>, Zewen Liu<sup>2</sup>, Jiujiang Zhou<sup>1</sup>, and Yajing Hua<sup>1</sup>

<sup>1</sup> College of Electrical and Information Engineering, Hunan University, Hunan, China  
`{wz6521, mao_jianxu}@126.com`

<sup>2</sup> College of Computer Science and Electronic Engineering, Hunan University,  
Hunan, China  
`1zewen106@163.com`

**Abstract.** In smoke and haze environment, images acquired by vision create serious distortion or degradation. Obtaining some inaccurate information from an unclear vision, it will have some bad impacts on outdoor activities. More and more common in recent years, the haze phenomena need to be further research. According to the images analysis of the atmospheric degradation model, this article puts forward the improved algorithm based on dark channel prior and morphology. Given the application of He's algorithm to defog, it makes brightness reduce. Therefore, the article firstly proposes to increase the brightness of image before processing, and then estimates the global atmospheric value, the initial transmission rate and the haze density using morphology method, finally substitutes into the simplified model to get the haze-free image. The experimental results show that the proposed algorithm can recover effectively and quickly degraded images. Meanwhile, this algorithm can keep the detail edges of images.

**Keywords:** Image defogging, dark channel, atmospheric light, morphology.

## 1 Introduction

Accompanied by the rapid development of intelligent transportation and machine vision, computer vision system has been widely applied in various fields, such as video surveillance system, road traffic driver assistance system, space cameras and medical equipment and so on. However, the current computer vision system has not yet been fully mature, so there are still some problems to be solved. When environmental factors are relatively poor, like fog, haze and other weather conditions, these images collected by computer vision system appear serious degradation, and thereby this phenomenon will have bad effects on the intelligent transportation system to obtain accurate information.

Currently, the image restoration methods are mainly the following two categories: physical and non-physical model approach [2]. Physical model approach is to explore the physical process of image degradation, and build their degradation model, then obtain the best estimate of the value of the image without fog to improve the quality of the image through solving the reverse process of lowering the quality. Non-physical

model approach is to ignore the physical causes of image degradation, whose main purpose is to correct the image color and enhance image contrast.

In recent years, haze removal algorithms for single image have been making significant progress. Relative to the foggy images, Tan [3] considers a haze-free image must have a higher contrast ratio compared with the input hazy image and he removes haze by maximizing the local contrast of the restored image. And then he uses the random (MRF) model to further normalize the results. The method can maximize the recovery images details and structure, which is applied in certain scenes to get better results. However, because this method disconnects from the physical model, and as to the higher saturation of the image itself, it is prone to distortion and color saturation easily. Furthermore, the results appear cavity defects in the area of the local depth discontinuity.

Fattal [4] using a simple model based on physical laws, proposes the method based on ICA. This method firstly assumes that the reflectance of the local small square is constant matrix. Secondly, it assumes that the surface reflectance and transmission in a small square is independent, and the reflectance direction can be estimated by ICA. Finally, using MRF model to infer the color of the entire image, the method can produce a clear and natural image and an effective depth chart. But given the limitations of the method model, heavy fog images cannot get a better processing result. Meanwhile, as to the method based on the color statistic, the method cannot apply for the gray-scale image, and usually it is not difficult to handle the heavy fog area without color.

Based on dark channel prior, He et al [5] proposes a dark channel prior to image foggy algorithm. This method presents a clear of image except sky region which has low intensity values at least one channel in RGB color channel. In hazy image, dark channel intensity values are mainly composed of air light. The method directly uses dark channel to estimate transmission map, and employs soft matting to refine it. The method for outdoor images has achieved good results, but for some light areas, the results will distort. In addition, the big question is that soft matting can consume large amounts of memory and computation time, which real-time requirements cannot be met. Thus, this method cannot be widely applied in computer vision system in practice.

Therefore, for the defects of the conventional defogging algorithms to single image, this article presents a fast algorithm based on dark channel prior and morphology. This article aims to guarantee a certain defogging effects, and tries one's best to reduce the complexity of the algorithm so that it can make defogging speed meet the requirements of real-time application system. Based on the analysis of atmospheric foggy image degradation model, the proposed method firstly enhances image brightness and contrast before restoring image, and then uses dark channel prior to obtain the global atmospheric light values of each channel through acquiring the coordinates of the largest and least piece of the original image, and the initial transmission rate, meanwhile estimates the haze density through morphology method, which can be effective to obtain the haze density and keep the edges of images. Finally the foggy images have been restored.

## 2 Atmospheric Degradation Model and Dark Channel Prior

### 2.1 Atmospheric Degradation Model

Scattering is the main factor of image degradation phenomena in harsh environments, such as fog, haze, smoke. In 1975, the atmospheric scattering model is proposed by McCartney, the formula is as follows:

$$H(x) = F(x)e^{-rd(x)} + A(1 - e^{-rd(x)}) \quad (1)$$

where  $x$  is the spatial coordinates of the image pixel,  $H(x)$  is the observed haze image (that is going to be defogged),  $F(x)$  is the resulting image after defogging,  $d(x)$  is the depth information of the object on coordinates  $x$ ,  $r$  is the atmosphere scattering coefficient,  $A$  is the global atmospheric light, usually,  $A$  is generally assumed to be a global constant and independent of the spatial coordinates.

The first part in formula (1) is called direct attenuation. Due to the effects of atmospheric particles scattering, a part of the object surface reflection of light lose because of scattering, scattering part not directly decays exponentially with the increasing of the propagation distance. The second part in formula (1) is called air light. With the increase of the propagation distance, the intensity of air light increases gradually.

Let  $t(x) = e^{-rd(x)}$ , the formula (1) can be further simplified as:

$$H(x) = F(x)t(x) + A(1 - t(x)) \quad (2)$$

where  $t(x)$  means the medium transmission rate describing the proportion between the light which is scattered and that which reaches the camera. Haze removal goal is to recover  $F(x)$  from  $H(x)$ .

### 2.2 Dark Channel Prior and Estimating the Global Atmospheric Light Rapidly

In the CVPR 2009 Conference, He et al [5] proposes the dark channel prior rule by statistics of the haze-free outdoor images. The rule can be described as: in a certain local area, some pixels value always at least one color channel is very low. In other words, the light intensity value of this region is very small. For an arbitrary image  $F(x)$ , the dark channel  $F^{dark}$  can be defined as:

$$F^{dark}(x) = \min_c (\min_{y \in \Omega(x)} F^c(y)) \quad (3)$$

where the superscript  $c$  represents three channels R, G, B.  $F^c$  is a color channel of  $F$  and  $\Omega(x)$  is a local patch of pixel  $x$ .  $F^{dark}$  is called the dark primary colors of  $F$ , it is low in most cases, and close to zero. The law is called dark primary colors priori through statistical observation.

Generally, in the defogging algorithm of most single image, the value of  $A$  is calculated from the pixel containing the fog. In reference [7], the value through artificial selected sky area is treated as the value of the global atmospheric light  $A$ . In reference [5], it takes 0.1% of the dark channel input image corresponding to the maximum brightness value as the value of the atmospheric light  $A$ . This method is reasonable which has a good result. But the process consumes time relatively. In reference [8] and [9], the magnitude and direction of the global atmospheric light are estimated by space geometry and optimization methods. This method is very complicated, and also requires a strong assumption, so it has a significant limitation. However, this article is to take the brightest pixel of image as the value of the global atmospheric light. This method has some limitations, but its results are still good after taking some experiments. It is most important that the quality can be guaranteed to defog while saving time, which accelerates the processing speed.

### 2.3 Estimating the Transmission Rate

For the calculation of the transmission rate, firstly it assumes that the transmission rate  $t(x)$  is known, the mean value has been estimated previously. Taking the minimum operation in the local patch on the haze imaging Equation (2), we have:

$$\min_{y \in \Omega(x)} H^c(y) = t(x) \min_{y \in \Omega(x)} F^c(y) + A^c(1-t(x)) \quad (4)$$

Notice that the minimum operation is performed on three color channels independently. This equation is equivalent to:

$$\min_{y \in \Omega(x)} \frac{H^c(y)}{A^c} = t(x) \min_{y \in \Omega(x)} \frac{F^c(y)}{A^c} + (1-t(x)) \quad (5)$$

And then we take the min operation among three color channels on the above equation and obtain:

$$\min_c \left( \min_{y \in \Omega(x)} \frac{H^c(y)}{A^c} \right) = t(x) \min_c \left( \min_{y \in \Omega(x)} \frac{F^c(y)}{A^c} \right) + (1-t(x)) \quad (6)$$

According to the dark channel prior, the dark channel value should tend to zero for haze free image, it means:

$$F^{dark}(x) = \min_c \left( \min_{y \in \Omega(x)} F^c(y) \right) \rightarrow 0 \quad (7)$$

As  $A^c$  is always positive, this leads to:

$$\min_c \left( \min_{y \in \Omega(x)} \frac{F^c(y)}{A^c} \right) \rightarrow 0 \quad (8)$$

From equation (8) and (6), we can obtain the transmission rate as follows:

$$t(x) = 1 - \min_c \left( \min_{y \in \Omega(x)} \frac{H^c(y)}{A^c} \right) \quad (9)$$

where  $H^c$  is a color channel of the fog image  $H$ .

For the transmittance rate, He et al [5] proposes a method of using soft matting to optimize it, but the computing time is longer.

In formula (9), image defogging thoroughly will make the image distortion, and so the depth information of the image will be lost. Therefore, a coefficient  $w$  ( $0 < w \leq 1$ ) is introduced in formula (9) in order to control the amount of the residual fog of defogging image. The finally transmittance rate is modified as:

$$t(x) = 1 - w \min_c \left( \min_{y \in \Omega(x)} \frac{H^c(y)}{A^c} \right) \quad (10)$$

After obtaining  $t(x)$  and  $A$ , the defogging image can be calculated according to equation (2) as follow:

$$F(x) = \frac{H(x) - A}{\max(t(x), t_0)} + A \quad (11)$$

where  $t_0$  is used to avoid the overly defogging. A typical value of  $t_0$  is 0.1.

## 2.4 Estimating the Haze Density Based on Morphological Filtering

It is the key for image defogging to estimate correctly the haze density, while the references [4] and [5] have one disadvantage of high time complexity which is difficult to achieve real-time defogging. Because morphological filtering method has some good effect on keeping image edges in the mutation edges of the scene. In this article, it effectively and rapidly estimates the haze density through morphological filtering instead of using Soft Matting to refine the transmission.

Inflation and corrosion are the two basic operation of mathematical morphology, which are dual operation.

If  $f(x)$  is the original image,  $g$  is the structural element, so the corrosion of image  $f(x)$  is defined as:

$$(f \odot g)(x) = \max\{y : g_x + y \ll f\} \quad (12)$$

Where max is the maximum operation. From the perspective of the geometric decay, the corrosion has shrinkage image effects.

Similarly, the inflation of image  $f(x)$  is defined as:

$$(f \oplus g)(x) = \min\{y : (g^\wedge)_x + y = f\} \quad (13)$$

where  $\min$  is the minimum operation. From the perspective of the geometric decay, the inflation has expansion image effects.

## 2.5 Specific Algorithm Flow

The specific algorithm implementation process of this method is shown below:

- 1) Increasing image brightness and calculating the global atmospheric light value;
- 2) Estimating the initial transmission rate through the results of the dark channel prior;
- 3) Refining the initial transmission rate by utilizing morphological opening operation;
- 4) Plugging the refined transmission rate  $t(x)$ , input haze image  $H(x)$  and the atmospheric light  $A$  in the formula (11) to obtain the haze-free image.

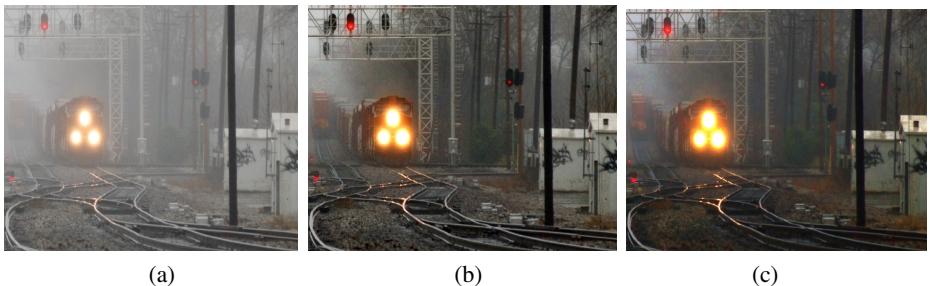
## 3 Experiment Results and Analysis

Experimental operating environment is a Windows7 operating system, the CPU of a dual core 2.94GHz, memory of 4GB, using Microsoft Visual C++ 6.0 to simulation algorithm. This article chooses reference [5] to compare and calculate the running time, image information entropy, contrast and mean value to evaluate the superiority of this algorithm.

Through taking the outdoor scene foggy images for experimental test, specific results are shown in the Fig.1, Fig.2, Fig.3. Select three images toys.jpg, train.bmp, forest.jpg, and their size is namely 500\*360, 600\*400, 1024\*768. From Fig.1, Fig.2, Fig.3 can be seen, there is no doubt that the processed images of He's and the method are both better than the original images from the naked eye. Carefully comparing the results in Fig.1, Fig.2, Fig.3, it can find that the proposed algorithm makes the image colors brighter, deeper.



**Fig. 1.** The experimental results of toys.jpg. (a) Original Image. (b) He's Result. (c) Improved Method's Result.



**Fig. 2.** The experimental results of train.bmp. (a) Original Image. (b) He's Result. (c) Improved Method's Result.



**Fig. 3.** The experimental results of forest.jpg. (a) Original Image. (b) He's Result. (c) Improved Method's Result.

### 3.1 Processing Time

If the algorithm puts into practice, it also needs to take the running time of the algorithm into account, and contrasting with the processing time can be measured to the difference of the computing complexity. The processing time of He's algorithm and the improved algorithm shows in Table 1. It is obvious that the improved algorithm's speed is more quickly than He's.

**Table 1.** The processing time of He's and the improved algorithm

Images	Size	Processing Time	
		He's Result/s	Improved Method's Result/s
Toys.jpg	500*360	2.1800	0.0366
Train.bmp	600*400	2.8600	0.0553
Forest.jpg	1024*768	9.4740	0.1394

### 3.2 Objective Evaluation Index

Currently, objective evaluation index used for degraded image, has primarily three types: full reference, half reference and no reference. Full reference and half reference require a clear image in the evaluation process. However, the article without clear images as reference images has no choice but to adopt no-reference method. Through a series of calculations, the article compares the quality indicators of original images with that of images after processing to evaluate the defogging effects. This article combines standard deviation, mean value and information entropy three aspects to weight on the defogging results. Generally speaking, in the image of each band, because of the haze what has high reflectivity, it increases the overall brightness of the image. If mean value drops over, the method has defogging effects; The size of standard deviation represents the images' resolution. The higher the value, the more the image sharpness; Information entropy reflects the amount of information. If entropy after defogging rises, it shows the image gets more information, and more clearness.

**Table 2.** Images of the objective evaluation index

toys.jpg	standard deviation	mean value	information entropy
Original Image	38.2991	163.0199	7.0593
He's Result	57.6456	103.7945	7.6194
Improved Method's Result	60.4188	118.0198	7.5980

(a) The Objective Evaluation Index of toys.jpg

train.bmp	standard deviation	mean value	information entropy
Original Image	31.1527	131.4778	6.8739
He's Result	40.7738	80.6517	7.1274
Improved Method's Result	43.1908	83.4126	7.2075

(b) The Objective Evaluation Index of train.bmp

forest.jpg	standard deviation	mean value	information entropy
Original Image	38.1409	118.4333	7.2452
He's Result	41.3120	95.0000	7.2145
Improved Method's Result	42.6026	77.4693	7.0759

(c) The Objective Evaluation Index of forest.jpg

As shown in Table 2 (a), (b), (c), comparing with the values from the table, the improved method reaches a defogging effect to a certain extent. From the datum in Table 2, it shows: standard deviations from the two methods are higher than original image, and it indicates the images after defogging become more clear, and this algorithm's standard deviation is higher than He's algorithm, so the improved method is more better than He's algorithm; Observation mean value can show that two methods have declined, which demonstrates the algorithm has some defogging effects; As for information entropy, the information entropy from two methods both increase some, relatively speaking, He's is more higher, it will detail some more. Overall, compared with He's method, these objective evaluation index are considerate, the improved method can reach a defogging effect, moreover enhance defogging speed.

## 4 Conclusion

Based on the atmospheric imaging optical model and dark channel prior, the article puts forward a simple fast algorithm of defogging images. By firstly increasing the brightness of an image before processing, images will get brighter after defogging. And then the article estimates the value of a global atmospheric light and the initial transmission rate  $t(x)$ , then optimizes the  $t(x)$  through morphology method, finally gets the defogging image. Experimental results show that the improved algorithm has a better robustness. And compared with He's method, this algorithm runs faster and objective assessment indicators are considerable.

**Acknowledgements.** This work is supported by National Natural Science Foundation of China (61072121, 60835004, 61271382), Hunan Provincial Natural Science Foundation of China (12JJ2035) and the Fundamental Research Funds for the Central Universities, Hunan university.

## References

1. Guo, J., Wang, X.-T., Hu, C.-P., Xu, X.: Image dehazing method based on neighborhood similarity dark channel prior. *Journal of Computer Applications* 5(31), 1224–1226 (2011)
2. Yu, J., Xu, D., Liao, Q.: Image defogging: a survey. *Journal of Image and Graphics* 16(9), 1561–1676 (2011)
3. Tan, R.T.: Visibility in Bad Weather from a Single Image. In: Proc of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Computer Society, Washington, DC (2008)
4. Fattal, R.: Single image dehazing. *ACM Transactions on Graphics*, 1–9 (2008)
5. He, K.-M., Sun, J., Tang, X.-O.: Single Image Haze Removal Using Dark Channel Prior. In: Proc of IEEE Conference on Vision and Pattern Recognition, pp. 1956–1963. IEEE Computer Society, Washington, DC (2009)
6. Tarel, J.P., Hautiere, N.: Fast Visibility Restoration from a Single Color Or Gray Level Image. In: Proc of IEEE International conference on Computer Vision. [S.I.], pp. 2201–2208. IEEE Press (2009)

7. Narasimhan, S.G., Nayar, S.K.: Interactive (De) Weathering of an Image Using Physical Models. In: Proceedings of the 2003 ICCV Vision Workshop on Color and Photometric Methods in Computer Vision, pp. 1387–1394. IEEE Press, Piscataway (2003)
8. Narasimhan, S.G., Nayar, S.K.: Vision and the Atmosphere. International Journal of Computer Vision 48(3), 233–254 (2002)
9. Narasimhan, S.G., Nayar, S.K.: Chromatic Framework For Visio. In: Bad Weather Proceedings of IEEE CVPR, pp. 598–605. IEEE, Washington, DC (2000)
10. Lv, X., Chen, W., Shen, I.-F.: Real-time Dehazing for Image and Video. In: 2010 18th Pacific Conference on Computer Graphics and Applications, pp. 62–69 (2010)
11. He, K., Sun, J., Tang, X.-O.: Guided Image Filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(35), 1397–1409 (2013)
12. Guo, F., Cai, Z.-X., Xie, B.: Video Defogging Algorithm Based on Fog Theory. Acta Electronica Sinica, 2019–2025 (2011)

# A New Image Structural Similarity Metric Based on K-L Transform

Cheng Jiang, Fen Xiao, and Xiaobo He

The College of Information Engineering, Xiangtan University, Xiangtan, Hunan, 41105, China  
Xiaof@xtu.edu.cn

**Abstract.** Recently, structural similarity image metric (SSIM) becomes the most popular model for image quality assessment (IQA). The idea behind SSIM is that natural images are highly structured, and estimate a general similarity of the image pairs from luminance, contrast and structure comparison. A novel similarity measure based on K-L transform is presented in this paper. It combines edge and texture components to provide a hierarchical description of image structure. We validate the performance of our algorithm with an extensive subjective study involving two sets of compressed images, the JPEG and the JPEG2000 images at the LIVE website. The experimental results show that the obtained quality metric had a high correlation with the subjective measure and outperforms SSIM.

**Keywords:** K-L transform, SSIM, human visual system, edge feature, texture feature.

## 1 Introduction

Image quality assessment plays an important role in numerous image applications and has attracted extensive attention. With the rapid development of digital imaging and communication technologies, images may be distorted during acquisition, compression, transmission and restoration. The visual quality of an image is quite important because most of applications are directed toward human users. And there are many automatic methods for quality monitoring assessment have been established.

The methods of image quality evaluation are mainly divided into subjective evaluation and objective evaluation. Subjective evaluation is generally regarded as a definitive method for assessing image quality and more consistent with human visual system (HVS). Nevertheless, it is time-consuming, sensitive to observer's background and motivation resources.

To deal with the limitations of the subjective method aforementioned, some early objective evaluations have been proposed, such as the mean squared error (MSE) and peak signal-to-noise (PSNR) [1]. The MSE and PSNR are appealing due to their simple numeration and clear physical meaning. But they are also widely criticized for not matching well with perceived visual quality [2]. In this instance, structural similarity (SSIM) has been proposed, which attempts to incorporate structural information into image comparison [3]. A class of metrics has been developed in both the space domain

and the transform domain, and tries to compare three components in different subbands. As the SSIM is failing in measuring the badly blurred images, Chen proposes a structural edge similarity metric (ESSIM) [4], which based on the assumption that the edge information is the crucial feature of image structure.

Remarkably, natural images consist of texture, structure and smooth regions. Perceived image distortion of any image strongly depends on the edge and texture feature [5]. The contributions of the two features to the structure similarity measure are dissimilar. The rate 2.3:1.68 between the two features is directly linked to the human psychology visual system [6]. Owing to the KLT (Karhunen-Loeve transform) is the most efficient transform in terms of energy compaction. KLT is used in the image block comparison. And the new metric considers both edge and texture features as the crucial structure feature, and combines them proportionally. In this paper, we use the Harris Response (HR) to check the edge pixels and the normalized Gray-level Co-occurrence Matrix will be used for texture feature extraction.

The remainder of this paper is organized as follows. Section 2 gives the knowledge of the new method including SSIM and the K-L transform. The proposed metric is described in Section 3. Section 4 presents the experimental results. Finally, section 5 draws the conclusion.

## 2 Background of the New Metric

### 2.1 The SSIM Index [2]

SSIM quantifies visual quality with a similarity measure between two patches  $x$  and  $y$ , which are the windows of the reference image and the destroyed image respectively, as the product of three components: luminance comparison, contrast comparison and structural comparison [4]. They can be defined as:

1. Luminance comparison:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

2. Contrast comparison:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

3. Structure comparison:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

where the  $\mu$  and  $\sigma$  correspond to the mean intensity and the standard deviation, respectively. The constants  $C_1$ ,  $C_2$  and  $C_3$  are included to avoid instability

where  $\mu_x^2 + \mu_y^2$ ,  $\sigma_x^2 + \sigma_y^2$  and  $\sigma_x \sigma_y$  are very close to zero, we choose  $C_1 = 0.01, C_2 = 0.03, C_3 = 0.03$  [2].

Then the SSIM value of the two blocks can be defined as:

$$SSIM = l(x, y)^\alpha c(x, y)^\beta s(x, y)^\gamma \quad (1)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are positive weights, typically set to 1.

The overall image quality can be evaluated by mean SSIM, which is defined as:

$$MSSIM = \frac{1}{M} \sum_{j=1}^M SSIM(x_j, y_j) \quad (2)$$

It is clear that the higher the value of  $MSSIM(x, y)$ , the more similar the images X and Y.

## 2.2 K-L Transform

It is obvious that the contrast measurement corresponds to the self-correlation calculation. In pursuit of contrast is equal to seek the correlation between the pixels. As the K-L space is the best one who has taken the strong correlation of pixels into account, we use it to obtain the contrast matrix.

Assume  $W = \{v_1, v_2, \dots, v_n\}$  is a nonnegative local image patch (for example  $5 \times 5$  image patch),  $v_i (i = 1, 2, \dots, n)$  denotes the column of digital image windows. Then for the image block W, the K-L transform is defined as:

$$F = \Phi^T (W - \mu) \quad (3)$$

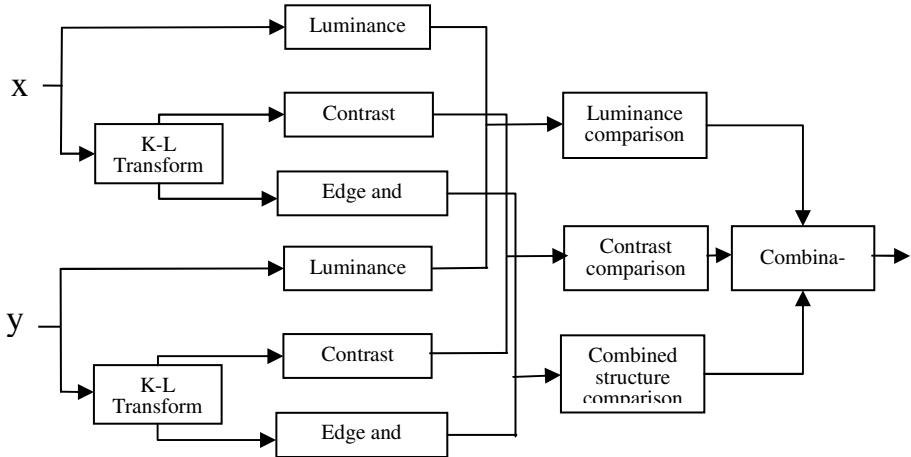
Where  $\mu = E(W) = \frac{1}{n} \sum_{i=1}^n v_i$  corresponds to the mean of  $W$ , and the column vector of  $\Phi$  is the eigenvector corresponding to the covariance matrix  $C_w$

$$C_w = E \left\{ (W - \mu)(W - \mu)^T \right\} \quad (4)$$

## 3 The K-L Based Image Structural Edge Similarity Metric

Among the SSIM components, the product of the variance and structure components perform nearly identically to the corresponding complete metric definition that uses all three components [7]. In other words, the performance of SSIM is basically depends on the performance of the variance and structure components. Consequently, if we want to conquer the limitation of feature extraction of SSIM, we'd better focus our mind

on the variance comparison and structure comparison. Consequently, we introduce the K-L transform to grasp the main character of an image and compute different structure feature sets, the edge feature and texture feature, proportionally. The system diagram of the proposed quality assessment system is shown in Fig.1.



**Fig. 1.** Diagram of the K-L based structure feature type similarity measurement (KLSSIM) system

### 3.1 The Image Components Comparison Based on K-L Transform

Before image components comparison, we firstly introduce K-L transform. The K-L transform can filter out the features that have no contribution to image quality assessment. So that we can use the K-L transform to get rid of the influence from useless features such as the flat.

The system introduces K-L transform before the contrast and structure comparisons. Suppose  $X$  and  $Y$  are two nonnegative image signals, one of them has perfect quality. We estimate the K-L based contrast measurement as the mean intensity of  $F$  calculated with eq. (3) and define the variance comparison:

$$kc(x, y) = \frac{2\mu_{F^x}\mu_{F^y} + C_4}{\mu_{F^x}^2 + \mu_{F^y}^2 + C_4} \quad (5)$$

where  $\mu_{F^x}$  and  $\mu_{F^y}$  correspond to mean intensity of  $F^x$  and  $F^y$  respectively,  $F^x$  and  $F^y$  are the output signals of K-L transform of distorted and reference images.

And the structure comparison includes two comparisons: edge and texture. We detect the edge points with Harris Response [8]. The Harris response is computed with three steps. First, each input patch  $W$  is filtered by a high pass filter to obtain

directional derivative images  $I_x$  and  $I_y$ . Second, the gradient information matrix  $C$  is computed to obtain the Harris response.

$$C = \begin{bmatrix} (I_x(x_i, y_i))^2 & (I_x(x_i, y_i)I_y(x_i, y_i)) \\ (I_x(x_i, y_i)I_y(x_i, y_i)) & (I_y(x_i, y_i))^2 \end{bmatrix} \quad (6)$$

Third, get the Harris Response  $R$  by the eigenvalues of matrix  $C$  from one point to another, the function written as:

$$R = \det(C) - k \cdot \text{tr}(C)^2 = \lambda_{\max} \lambda_{\min} - k (\lambda_{\max} + \lambda_{\min})^2 \quad (7)$$

where  $k$  denotes a constant set to 0.06 in our experiments,  $\det(\cdot)$  and  $\text{tr}(\cdot)$  are determinant and trace of a matrix respectively. If  $R$  is negative (positive), the pixel is a corner (edge) pixel [9] [10]. All of the negative or positive pixels make up of our edge feature. Similarly, we define the edge comparison as

$$es(x, y) = \frac{\sigma_{R_{F^x} R_{F^y}} + C_5}{\sigma_{R_{F^x}} \sigma_{R_{F^y}} + C_5} \quad (8)$$

Where  $R_{F^x}$  and  $R_{F^y}$  are the Harris Response  $R$  of  $F^x$  and  $F^y$  respectively.

The texture feature matrix is generated by the Gray-level Co-occurrence Matrix (GLCM), as its good performance on texture extraction. In general terms, the computational directions are  $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ , and the parameters are the energy, entropy, inertia and correlation. We use the mean and stander deviation of the four parameters as the final eight-dimensional texture feature  $P = \{P_1, P_2, \dots, P_8\}$ . The details on the process of texture calculation can be found in [11]. And we get the texture comparison as

$$ts(x, y) = \frac{1}{8} \sum_{i=1}^8 T_i(F^x, F^y) \quad (9)$$

Where  $T_i(F^x, F^y) = \frac{\sigma_{P_{iF^x} P_{iF^y}} + C_6}{\sigma_{P_{iF^x}} \sigma_{P_{iF^y}} + C_6}, 1 \leq i \leq 8$ ,  $P_{iF^x}$  and  $P_{iF^y}$  are the texture vectors obtained by the Gray-level Co-occurrence Matrix of  $F^x$  and  $F^y$  respectively.

### 3.2 KLSSIM

The new metric KLSSIM can be defined as follows:

$$KLSSIM = l(x, y)kc(x, y)ps(x, y) \quad (10)$$

Where  $l(x, y)$  denotes the luminance comparison,  $kc(x, y)$  is variance comparison defined in eq. (5), and  $ps(x, y)$  denotes the combined structure comparison

$$ps(x, y) = 0.575es(x, y) + 0.42ts(x, y) \quad (11)$$

where the combined ratio is directly linked to the human perception, and the edge comparison  $es(x, y)$  and texture comparison  $ts(x, y)$  are estimated by the eq. (8) and (9).

Accordingly, the overall image quality can be evaluated by mean KLSSIM, which is defined as:

$$MKLSSIM = \frac{1}{M} \sum_{j=1}^M KSFTSIM(x_j, y_j) \quad (12)$$

## 4 Experimental Results and Discussion

To evaluate the performance of the proposed KLSSIM, we use 223JPEG and 221JPEG2000 compressed images from LIVE data [12]. The bit rates are from 0.150 to 3.336 and 0.028 to 3.150 bits/pixel. Every JPEG and JPEG2000 image are marked by 13~20 and 25 observers, who are mostly male college students [2]. An ideal image quality assessment (IQA) should mimic the human observer. And the Mean Opinion Score (MOS) [13] is the well-known criteria to evaluate whether an IQA is corresponding to the human perception. So in order to check the performance of KPSSIM, we should figure out the distance between MOS and the new IQA value.

An image, which is compressed with low compression ration, has low subjective score as it loses lots of information and is destroyed badly. On the contrary, an image with high compression ration will receive a high subjective score as it just loses a little information. Fig. 2 and Fig. 3 show some example images that compressed to different bits. It is obvious that when the image has low bits its KLSSIM value will be low, and when the compression ratio is high, the image will get a high KLSSIM score. As can be seen from Fig. 2 and Fig. 3, the proposed KLSSIM is corresponding to the human perception of the compressed images with different distortion ratio.

Compared with SSIM, KLSSIM is better at assessing the bad blurred images. Fig. 4 shows some example images with different distortion. From Fig. 4 we can see that image (b) and image (c) have different degree of degradation while they have the almost same MSSIM value. It is obvious that image(c) is worse than image (b), so we wish the quality score of image(c) should be lower. Actually, in the result, KLSSIM can show this quality distance well.



**Fig. 2.** JPEG2000 images comparison. The original (a) "sauling4," (b)"buildings," and (c)"house" images. (d) Compressed to 0.59656 bits/pixel, PSNR=55.547dB, MKLSSIM=0.65507. Image (e) compressed to 0.1264 bits/pixel, PSNR=25.847dB, MKLSSIM= 0.33594. Image (f) compressed to 0.082499 bits/pixel, PSNR=39.032dB, KPMSSIM=0.2862.

In order to provide quantitative measures on the performance of the objective quality assessment model, we use in terms of the performance measures such as the logistic function, Person correlation coefficient (CC) between the objective/subjective scores,

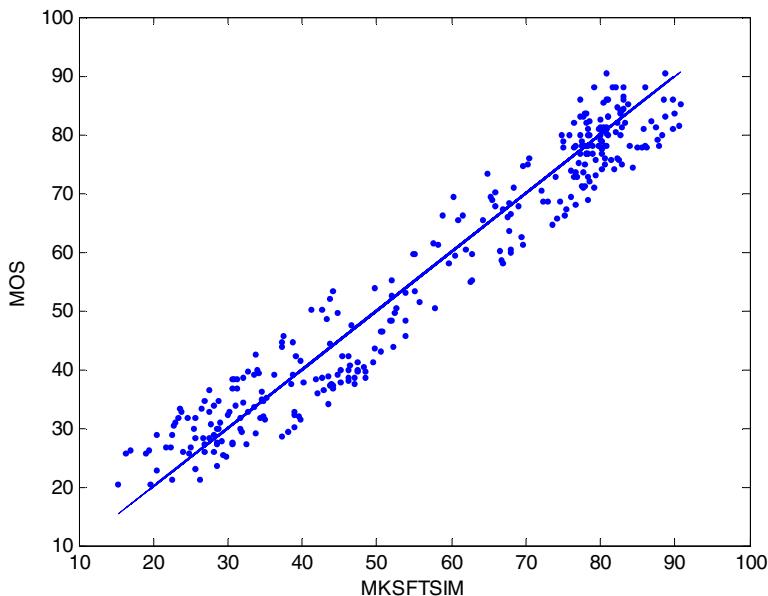


**Fig. 3.** JPEG images Comparison. The original (a)"caps," (b)"monarch," and (c)"ocean" images. Image (d) compressed to 0.85118 bits/pixel, PSNR=78.569dB, MKLSSIM =0.7601. Image (e) compressed to 0.32107bits/pixel, PSNR= 9.16dB, MKLSSIM = 0.44662. Image (f) compressed to 0.15037bits/pixel, 28.675dB, MKLSSIM=0.1613.

mean absolute error (MAE) and root mean squared error (RMS), which are employed in the video quality expert group (VQEG) Phase I FR-TV test [14]. Logistic function is used in a fitting procedure to provide a nonlinear mapping between the objective/subjective scores [2]. CC reflects the degree of approach between the objective/subjective scores. A good performance IQM should have a high CC numerical value. We calculate the MAE and RMS after nonlinear regression, both of the two metrics have the capability of showing the error ratio between the objective/subjective scores. More details on these metrics can be found in [14]. The evaluation results for comparison with other IQM are given in Table1 and Table2. From the two tables we can consider a conclusion that KPSSIM performs better than all of the other comparable models.



**Fig. 4.** Comparison of "stream" image with different distortion, (a) original image. (b) Gaussian blur image. MSE=89.624, MSSIM=0.3836, MKLSSIM=0.3927. (c) White noise image. MSE=95.837, MSSIM=0.3999, MKLSSIM=0.2898.



**Fig. 5.** Scatter plots of subjective mean score versus model prediction

**Table 1.** Quality assessment results on JPEG compressed images

Method	CC	MAE	RMS
PSNR	0.899	10.05	11.75
MSSIM	0.912	8.27	9.73
MKLSSIM	0.943	7.15	8.78

**Table 2.** Quality assessment results on JPEG2000 compressed images

Method	CC	MAE	RMS
PSNR	0.897	10.14	12.23
MSSIM	0.906	8.90	11.90
MKLSSIM	0.939	6.65	8.72

## 5 Conclusions

In this paper, we proposed an effective IQM which is KL-based structural features similarity, called KLSSIM. The new IQM measures the contrast and the structural comparison based on the K-L transform. According to human perception we combine the edge feature comparison with the texture feature comparison proportionally. And consider the combine as the structural comparison. For detecting the edge feature, we adopt the Harris Response in which the texture feature can be extracted by the Gray-level Co-occurrence Matrix. Experimental results show that the proposed KLSSIM is a better IQM than other IQM that have been compared in this paper. Further research will focus on the extension of the KLSSIM to video quality assessment and other image processing.

## References

1. Winkler, S., Mohandas, P.: The evolution of video quality measurement: from PSNR to hybrid metrics. *IEEE Transactions on Broadcasting* 54(3), 660–668 (2008)
2. Wang, Z.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)
3. Zujovic, J., Pappas, T.N., Neuhoff, D.L.: Structural similarity metrics for texture analysis and retrieval. *IEEE Transactions on Image Processing* 22(7), 2545–2558 (2013)
4. Chen, G.-H., Yang, C.-L.: Edge-based structural similarity for image quality assessment. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal, May 14-19, pp. 933–936 (2006)
5. Zhao, X., et al.: Structural texture similarity metrics for retrieval applications. In: Proc. of the 15th IEEE International Conference on Image Processing, October 12-15, pp. 1196–1199 (2008)

6. Li, J., Chen, Z., Lu, C.: A three-step objective image quality assessment method. In: Proc. of the International Conference on Virtual Reality and its Application in Industry, April 9-12, pp. 526–531 (2002)
7. Rouse, D.M., Sheila, S.: Understanding and simplifying the structural similarity metric. In: Proc. of the 15th IEEE International Conference on Image Processing, October 12-15, pp. 1188–1191 (2008)
8. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: Proceeding 4th Alvey Vision Conference, Manchester, August 31, pp. 147–151 (1988)
9. Kim, D.-O., Han, H.-S.: Gradient Information-Based Image Quality Metric. IEEE Transactions on Consumer Electronics 56(2), 930–936 (2010)
10. Kim, D.-O., Park, R.-H.: New Image Quality Metric Using the Harris Response. IEEE Signal Processing Letters 16(7), 616–619 (2009)
11. Modestino, J.A., Zhang, J.: A markov random field model based approach to image interpretation. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(6), 606–615 (1992)
12. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing 15(11), 3441–3452 (2006)
13. Wang, Z., Bovik, A.C.: Modern Image Quality Assessment. Morgan & Claypool, New York (2006)
14. VQEG, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment (March 2000), <http://www.vqeg.org/>

# A New Restoration Algorithm for Single Image Defogging

Fan Guo, Hui Peng, and Jin Tang

School of Information Science and Engineering, Central South University, Changsha, China  
guofancsu@163.com

**Abstract.** Fog is an atmospheric phenomenon that significantly degrades the visibility of outdoor scenes. Thus, this paper presents an algorithm to remove fog for a single image. The method estimates the transmission map of image degradation model by assigning labels with MRF model and optimizes the map estimation process using the graph-cut based  $\alpha$ -expansion technique. The algorithm goes with two steps: first, the transmission map is estimated using a dedicated MRF model combined with the bilateral filter. Then, the restored image is obtained by taking the estimated transmission map and the airlight into the image degradation model to recover the scene radiance. A comparative study is proposed with a few other state of the art algorithms which demonstrate that better quality results can be obtained using the proposed method.

**Keywords:** image, restoration, fog removal, image degradation model, transmission map.

## 1 Introduction

The quality of photograph in our daily life is easily undermined by the aerosols suspended in the medium, such as dust, mist, or fumes. This has an effect on the image, e.g., contrasts are reduced and the surface color becomes faint. Such degraded photographs often lack visual vividness and offer a poor visibility of the scene contents. The goal of defogging algorithms is to enhance and recover the detail of the scene from foggy image. There are many circumstances that accurate fog removal algorithms are needed. In computer vision, most automatic systems for surveillance, intelligent vehicles, object recognition, etc., assume that the input images have clear visibility. However, this is not always true in bad weather. Therefore, removing fog from a single image is very important and useful. Since the process of image defogging depends on the depth of the scene, thus the essential problem that must be solved in most image defogging methods is scene depth estimation. This is not trivial and requires prior knowledge. In this paper, the transmission map is estimated by assigning labels with MRF model and optimizes the map estimation process using the graph-cut based  $\alpha$ -expansion technique. Experimental results show that good defogging effect may be produced by using the proposed method.

Since the importance of the defogging algorithm, many defogging works have been done. Graphical model (GM), as a probabilistic model combined probability with graph, is an important way to solve this problem. The models can be divided into two

types: directed graph and undirected graph. Generally, a directed GM is a Bayesian network (BN) when the graph is acyclic, meaning there are no loops in the directed graph. The relationships in a BN can be described by local conditional probabilities [1]. In [2, 3], a Bayesian defogging method that jointly estimates the scene albedo and depth from a single foggy image is introduced by leveraging their latent statistic structure. The undirected graph refers to Markov Random Field (MRF). Since MRF is undirected and may be cyclic, it can represent certain dependencies that a BN cannot, which providing a new way for image defogging due to the dependencies exist between the neighboring pixels. The defogging algorithm in [4] based on the observation that the surface Lambertian shading factor and the scene transmission are locally independent in order to separate the fog from the scene, and then a Gaussian-Markov random field is used to smooth the transmission values. In [5], a cost function in the framework of MRF is developed to enhance the visibility of the images. However, the results obtained by this method tend to have larger saturation values than those in the actual clear-day images. In [6], the scene geometry and the alpha-expansion optimization technique are employed to improve the robustness of single image dehazing algorithm. Recently, image defogging based on MRF model has made significant progresses [7-8]. In [7], the image defogging problem is decomposed into two steps: first infer the atmospheric veil using a dedicated MRF model, and second estimate the restored image by minimizing MRF energy. In this MRF model, the flat road assumption is introduced to achieve better results on road images. In [8], a MRF model of both stereo reconstruction and defogging problems is combined into a unified MRF model to take advantages of both stereo and atmospheric veil depth cues. Thus, the stereo reconstruction and image defogging in daytime fog can be solved using the new MRF model. In [9], a multi-level depth estimation method based on MRF model is presented for image defogging. The method integrated the characteristic of dark channel prior into the MRF model to estimate an accurate depth map. MRF is applied here to label the depth level in adjacent region for the compensation of wrong estimated regions. The textures in the scene are the critical element served as the smooth term in the MRF model. These fog removal algorithms are the most representative of MRF defogging methods. However, the color or the profile of the scene objects may sometimes look unnatural for the defogged results.

## 2 Background

### 2.1 Markov Random Field

Many vision problems can be naturally solved by using MRF technique. Markov random field theory is a branch of probability theory for analyzing the spatial or contextual dependencies of physical phenomena. It is often used in visual labeling to establish probabilistic distribution of interacting labels. Here, we use MRF to estimate the transmission map in image degradation model. It is an undirected graph, and adjacent nodes are connected to determine the depth of real scene [9]. We associate hidden layer with the dense level of fog and observation layer with the initial

transmission map, and then a MRF model is provided to a cost function as shown in follows:

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} V_{p,q}(f_p, f_q) , \quad (1)$$

In (1),  $f = \{f_p \mid p \in P\}$  is a labeling of image  $P$ ,  $f_p$  is the label of pixel  $p$  in the image  $P$ , and  $f_p = \{1, 2, 3, \dots, k\}$ .  $q$  is the neighbor of  $p$ ,  $N$  is the set of pairs of pixels defined over the standard four-connection neighborhood.  $E(f)$  is for minimizing sums of two types of terms. The first term  $D_p(\cdot)$  is a data function. The smaller difference between the pixel and its label, the smaller  $D_p(\cdot)$  will be.  $D_p(\cdot)$  penalizes a label  $f_p$  to pixel  $p$  if it is too different with the observed data  $I_p$ . Generally, the data term in MRF model or other energy function of regularization based optimization problem constructs the constraint between the expected variable and some known observations about the variable. The second term  $V_{p,q}(\cdot)$  is a smooth function (or called discontinuity-preserving) [10, 11]. The smaller difference among the labels of pixels in set  $N$ , the smaller  $V_{p,q}(\cdot)$  will be.  $V_{p,q}(\cdot)$  encourage the integrity of an image by penalizing two neighboring labels  $f_p$  and  $f_q$  if they are too different. The choice of  $V_{p,q}(\cdot)$  is a critical issue, and in the proposed defogging method we apply the geometry prior to obtain this term. With the smoothing term, the saturated colors at each pixel can be computed with a reasonable smoothing. Thus, for the transmission map estimation, the data function represents the probability of pixel  $p$  having transmission association with label  $f_p$ . The smooth function encodes the probability where neighboring pixels should have similar depth. For the transmission map estimated by using MRF model, the small value of the label on behalf of the deeper depth in the scene, vice versa. The relabeling results would be the initial transmission map of the proposed method. However, there still exists some redundant details need to be removed.

## 2.2 Geometry Prior for Foggy Image

In this section, we'll present the geometry prior that used in the transmission map estimation of the proposed algorithm. Light passing through a scattering medium is attenuated and distributed to other directions. This can happen anywhere along the path, and leads to a combination of radiances incident towards the camera. Formally, to express the relative portion of light that managed to survive the entire path between the observer and a surface point in the scene, the defined transmission map  $t_i$  combines the geometric distance  $d_i$  and medium extinction coefficient  $\beta$  (the net loss from scattering and absorption) into a single variable [12]:  $t_i = e^{-\beta d_i}$ . Thus, the following geometry prior is reasonable: assuming that  $\beta$  is constant over the image, the variations of transmission are due to the distance  $d$  between the scene point and the camera, and the larger distance means the smaller intensity in the transmission map. For most outdoor images, especially the surveillance images, the transmission map can be expressed in the component of distance along the ground and height above the ground. As long as the scene does not contain any cave-like surfaces, such as the space underneath a bridge, the distance along the ground to the visible scene point is a monotonically increasing function of image plane height which starts from

the bottom of image to the top. Therefore, the object which appears closer to the top of the image is usually further away. To verify the effectiveness of the geometry prior, we collect an outdoor image sets from internet and real captured photos. Statistical results show that about 83% images support our geometry prior. Since the geometry prior is a kind of statistic, it may not work for some particular images when the scene objects which near observer appear on the top of images, e.g., twigs, walls, trees, etc. the geometry prior is invalid. Fortunately, the proposed MRF-based method can still produce a reasonable transmission map without creating significant errors in the restored image.

### 3 The Proposed Algorithm

Specifically, the proposed algorithm has three steps to remove fog from a single image: the first one is computing the airlight according to the three distinctive features of sky region. The second step is computing the transmission map with the MRF model and the bilateral filter. The goal of this step is assigning the accurate pixel label using the graph-cut based  $\alpha$ -expansion and removing the redundant details using the bilateral filter. Finally, with the estimated airlight and transmission map, the scene radiance can be recovered according to the image degradation model.

#### 3.1 Airlight and Transmission Map Estimation

The presence of aerosols in the lower atmosphere means that the light may scatter and be absorbed while traveling through the medium [13]. This can happen anywhere along the path, and lead to a combination of radiances incident towards the camera. The image degradation model that widely used to describe the formation of the foggy image is as follows [2]:

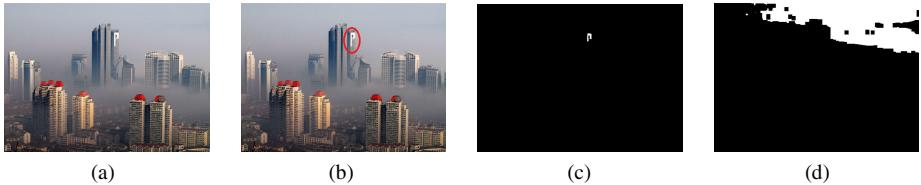
$$I(\mathbf{x}) = J(\mathbf{x})t(\mathbf{x}) + A(1-t(\mathbf{x})) \quad (2)$$

where  $I(\mathbf{x})$  is the observed intensity corresponding to the pixel  $\mathbf{x}=(x, y)$  and also the input foggy image,  $J(\mathbf{x})$  is the scene radiance and also the fog removal image,  $A$  is the airlight, and  $t(\mathbf{x})$  is the transmission map, which is the key factor for image defogging. In (2), the first term  $J(\mathbf{x})t(\mathbf{x})$  is called direct attenuation model, and the second term  $A(1-t(\mathbf{x}))$  is called airlight model. Theoretically, the goal of fog removal is to recover  $J(\mathbf{x})$  from the estimated  $A$ ,  $t(\mathbf{x})$  and the original image  $I(\mathbf{x})$ .

##### 3.1.1 Airlight Estimation

To estimate the airlight, we sum up the three distinctive features of sky region according to the nature fact of a large quantity of image sky regions. The distinctive features of sky region are: (i) bright minimal dark channel, (ii) flat intensity, and (iii) upper position. For the first feature, the pixels that belong to the sky region should satisfy  $I_{min}(\mathbf{x}) > T_v$ , where  $I_{min}(\mathbf{x})$  is the dark channel and  $T_v$  is the 95% of the maximum value of  $I_{min}(\mathbf{x})$ . For the second feature, the pixels should satisfy the constraint  $N_{edge}(\mathbf{x}) < T_p$

where  $N_{edge}(\mathbf{x})$  is the edge ratio map and  $T_p$  is the flatness threshold. Due to the third feature, the sky region can be determined by searching for the first connected component from top to bottom. Thus, the atmospheric light  $A$  is estimated as the maximum value of the corresponding region in the foggy image  $I(\mathbf{x})$ . Fig. 1 shows the contrastive airlight estimation result for comparing the performance of the proposed sky region detection with the “brightest pixel” method. From Fig. 1, we can see that the proposed technique is more robust than the “brightest pixel” method. Notice that the proposed technique can gracefully handle the input foggy images even without sky regions by using the image degradation and MRF model. If not so, other methods, such as contrast enhancement, can be used to remove fog from input images.



**Fig. 1.** Airlight estimation example. (a) Input image. (b) The brightest pixels region. (c) The wrong sky region obtained by the “brightest pixel” method. (d) The correct sky region obtained by the proposed method.

### 3.1.2 Initial Transmission Map Estimation

Transmission map estimation is the most important step for image defogging, here we use the graph-cut based  $\alpha$ -expansion method to estimate the map  $t(\mathbf{x})$ , as it is able to handle regularization and optimization problem, and has a good track record with vision-specific energy function [14]. Specifically, each elements  $t_i$  of the transmission map is associated with a label  $x_i$ , where the set of Labels  $L=\{0, 1, 2, \dots, l\}$  represents the transmission values  $\{0, 1/l, 2/l, \dots, 1\}$ . Before labeling, we first convert input RGB image to gray-level image. Thus, the number of Labels is 32 since the labeling unit of pixel value is set to be 8 and  $l=31$ . The most probable labeling  $x^*$  minimizes the associated energy function:

$$E(x) = \sum_{i \in P} E_i(x_i) + \sum_{(i, j) \in N} E_{ij}(x_i, x_j) \quad (3)$$

where  $P$  is the set of pixels in the unknown transmission  $t$ , and  $N$  is the set of pairs of pixels defined over the standard four-connect neighborhood. The unary function  $E_i(x_i)$  is the data term representing the possibility of pixel  $i$  having transmission  $t_i$  associated with label  $x_i$ . The smooth term  $E_{ij}(x_i, x_j)$  encodes the possibility where neighboring pixels should have similar depth.

For data function  $E_i(x_i)$ , which represents the possibility of pixel  $i$  having transmission  $t_i$  associated with label  $x_i$ , we first convert the input RGB image  $I_i$  to gray-level image  $I'_i$ , and then compute the absolute differences between each pixel value and the label value. The process can be written as:

$$E_i(x_i) = |I_i \times \omega - L(x_i)| \quad (4)$$

In (4),  $I_i$  is the intensity of pixel in the gray-level image ( $0 \leq I_i \leq 1$ ).  $L(x_i)$  is the each element in the set of Labels  $L=\{0, 1/l, 2/l, \dots, 1\}$ . The parameter  $\omega$  is introduced to ensure that  $I_i$  and  $L(x_i)$  have same order of magnitude. Since each pixel value of our initial transmission map is expressed by the label  $x_i$ , and the labels depend on the pixel intensity of gray-level image. Thus, the data function can model two different physical quantities together, i.e., pixel intensity and transmission.

The smooth function  $E_{ij}(x_i, x_j)$  encodes the possibility where neighboring pixels should have similar depth. Inspired by the work [6], we use the linear cost function, which is solved by  $\alpha$ -expansion:

$$E_{ij}(x_i, x_j) = w|x_i - x_j| \quad (5)$$

From the geometry prior, we know that objects which appear closer to the top of the image are usually further away. Thus, if we consider two pixels  $i$  and  $j$ , where  $j$  is directly above  $i$ , we have  $d_j > d_i$  according to the geometry prior. Thus, we can deduce that the transmission  $t_j$  of pixel  $j$  must be less than or equal to the transmission  $t_i$  of pixel  $i$ , that is  $x_j \leq x_i$ . For any pair of labels which violate this trend, a cost  $c > 0$  can be assign to punish this pattern. Thus, the smooth function in Eq. (5) can be written as:

$$E_{ij}(x_i, x_j) = \begin{cases} c & \text{if } x_i < x_j, \\ w|x_i - x_j| & \text{otherwise.} \end{cases} \quad (6)$$

The parameters  $w$  and  $c$  are used to control the aspect of defogging effect. The value of  $w$  controls the strength of the detail enhancement, and is usually set to 0.01. The cost  $c$  controls the strength of the color recovery, and is usually set to 100. The two parameters are useful to compromise between highly enhanced details where colors may appear too dark, and less restored details where colors are brighter. Besides, the weights associated with the graph edge should be determined. If the intensity of two neighboring pixels in the input foggy image  $I$  are less than 15 in each channel, which means the two pixels have high possibility of sharing the same transmission value. Thus, the cost of the labeling is increased by  $15\times$  to minimize the artifacts due to the depth discontinuities in this case. Taking the data function and the smooth function into the energy function equation (3), the pixel label of transmission map can be estimated by using the graph-cut based  $\alpha$ -expansion. In our method, the gco-v3.0 library [14] developed by Veksler *et al.* is adopted for optimizing multi-label energies via the  $\alpha$ -expansion. It supports energies with any combination of unary, pairwise, and label cost terms [15, 16]. Thus, we use the library to estimate each pixel label in initial transmission map. By using the functions defined in the optimization library, we can obtain each pixel label  $x_i$ . Then, a proper intensity value of the initial transmission map can be assigned to each image pixel. Specifically, for each label  $x_i$ , we have

$$t_{ini}(\mathbf{x}) = 255 - (x_i - 1) \times 8 \quad (7)$$

In Eq. (7),  $t_{ini}$  is the initial transmission map estimated by the proposed MRF-based algorithm. An illustrative example is shown in Fig. 2. In the figure, Fig. 2(b) shows the initial transmission map estimated using the algorithm presented above, its corresponding restored result is shown in Fig. 2(c). One can clearly see that the appearance of the scene objects in the restored image looks one-dimensional.



**Fig. 2.** True example. (a) Input image. (b) Initial transmission map. (c) Restored result obtained using (b). (d) Bilateral filter to (b). (e) Restored result obtained using (d).

### 3.1.3 Refined Transmission Map Estimation

As shown in Fig. 2, there is obvious deficiency in the recovered image in the discontinuities of the transmission map obtained by MRF model. For example, the red bricks and the slots between them should have the same depth values. However, as shown in Fig. 2(b), one can clearly see the slots between the bricks in transmission map estimated by the MRF-based algorithm. In order to handle these discontinuities, many works adopt the bilateral filter to refine the transmission map estimation. Thus, the redundant details of the transmission map  $t_{ini}$  estimated by the algorithm presented above can be effectively removed, which improves the restored result with better detail enhancement capability. This process can be written as:

$$t(\mathbf{u}) = \frac{\sum_{\mathbf{p} \in N(\mathbf{u})} W_c(\|\mathbf{p} - \mathbf{u}\|) W_s(|t_{ini}(\mathbf{u}) - t_{ini}(\mathbf{p})|) t_{ini}(\mathbf{p})}{\sum_{\mathbf{p} \in N(\mathbf{u})} W_c(\|\mathbf{p} - \mathbf{u}\|) W_s(|t_{ini}(\mathbf{u}) - t_{ini}(\mathbf{p})|)} \quad (8)$$

where  $t_{ini}(\mathbf{u})$  is the initial transmission map corresponding to the pixel  $\mathbf{u}=(x, y)$ ,  $N(\mathbf{u})$  is the neighbors of  $\mathbf{u}$ . The spatial domain similarity function  $W_c(x)$  is a Gaussian filter with the standard deviation  $\sigma_c$ :  $W_c(x) = e^{-x^2/2\sigma_c^2}$ , and the intensity similarity function  $W_s(x)$  is a Gaussian filter with the standard deviation  $\sigma_s$ , it can be defined as:  $W_s(x) = e^{-x^2/2\sigma_s^2}$ . In our experiments, the value of  $\sigma_c$  and  $\sigma_s$  is set as 3 and 0.4, respectively. Thus, we can obtain the final refined transmission map, as shown in Fig. 2(d), and Fig. 2(e) is the restored result obtained using the refined map. From Fig. 2(e), one can see that the restored result obtained using the bilateral filter has more layer and stereoscopic feelings compared with the result [see Fig. 2(c)] obtained without using the filter.

### 3.2 Scene Radiance Recovery

Since now we already know the input haze image  $I(\mathbf{x})$ , the final refined transmission map  $t(\mathbf{x})$  and the airlight  $A$ , we can obtain the final fog removal image  $J(\mathbf{x})$  according to the image degradation model. The final defogging result  $J(\mathbf{x})$  is recovered by:

$$J(\mathbf{x}) = \frac{I(\mathbf{x}) - A}{\max(t(\mathbf{x}), t_0)} + A \quad (9)$$

where  $t_0$  is application-based, and it is used to adjust the fog kept at only the farthest reaches of the image. If the value of  $t_0$  is too large, the result has little defogging effect, and if the value is too small, the color of fog removal result seems oversaturated. Experiment shows that when  $t_0$  is set to be 0.2, we can get visual pleasing results in most cases.

## 4 Experimental Results

To evaluate the performance of various defogging algorithms, we compared our defogging algorithm with a few other state of the art algorithms. The first image of Fig. 3(b) shows the defogging result obtained by Fattal [17]. As can be seen in the figure, Fattal's method can produce a visual pleasing result. However, the method is based on statistics and requires sufficient color information and variance. If the fog is dense, the color information used in that method is not enough to reliably estimate the transmission. Then, we compare our method with Tan's work [18] in the second images of Figs. 3(b) and 3(d). The colors of Tan's result may sometimes over saturate or distort. For example, the color of the sky and road region in the Tan's result is turned into yellow, as shown in the figure. We also give He's work [19] in the third images of Fig. 3(b). He's algorithm can achieve a good enhancement effect for most outdoor images. However, when the scene objects are inherently similar with the airlight, the dark channel prior used in He's method will be invalid. In this case, the defogging result of He's algorithm is not visual pleasing, as shown in Fig. 3(b). The fourth images of Figs. 3(b) and 3(d) show a comparison between results obtained by Carr [6] and our algorithm. It can be seen that our algorithm tend to enhance details better than Carr's result, and the color of our result seems more close to the original input image. The fifth and sixth images of Figs. 3(b) and 3(d) show the results of our method and Caraffa's methods [7, 8]. From these images, we can see that although the results we get can't thoroughly remove the fog in very dense fog regions compared with Caraffa's methods, such as the buildings and the trees far away, our results appear natural in both color and the profile of the scene objects. The seventh images of Figs. 3(b) and 3(d) show a comparison between results obtained by Wang [9] and our method. One can clearly see that the color of the sky region in Wang's result seems a little inconsistent with that of the original foggy image. Experiments on a large quantity of outdoor images confirm the above conclusions. Since transmission map is very important to recover a good result, we also present the estimated transmission maps in Fig. 3(c).



**Fig. 3.** The comparison between recent fog removal work. (a) Original foggy images. (b) Corresponding fog removal results obtained by recent defogging algorithms. From left to right: Fattal's, Tan's, He's, Carr's, Caraffa's and Wang's results. (c) Our estimated transmission map. (d) Our fog removal results.

Therefore, results on a variety of haze or fog images show that the defogging image obtained with our algorithm seems visually close to the result obtained by Fattal, He and Carr, with better color fidelity and less halo artifacts compared with Tan, Caraffa and Wang. For our proposed algorithm, it takes about 2 minute to process a 600×400 pixel image. Notice that when the image size is small, the proposed method has a relatively faster speed. For example, only 3 second is needed to process a 250×190 pixel image using the proposed method. All the algorithms are tested by executing MATLAB on a PC with 3.00GHz Intel Pentium Dual-Core Processor. The speed can be further improved by using efficient parallel computation with a GPU.

## 5 Conclusion

Image defogging is an important issue in computer vision. In this paper, a new defogging algorithm was presented based on MRF model. The problem was formulated as estimation of transmission map with  $\alpha$ -expansion optimization. The algorithm is implemented by two steps: the transmission map is first estimated using a dedicated MRF model and the bilateral filter. Once the map is inferred, the restored image can be obtained according to the image degradation model. Experimental results demonstrate that the proposed algorithm can produce visually pleasing defogging results and tend to enhance the image contrast, which is better than previous techniques. However, the color of our defogging results sometimes seemed over-saturated. Nevertheless, we could improve the overall quality of a foggy image by enhancing the main details, and the algorithm could be further improved by employing better prior for the data and smooth function of the MRF model. In the future, we intend to investigate the case of various kinds of fog and speed up the proposed algorithm for real-time processing.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China (71271215, 71221061, and 91220301), the International Science & Technology Cooperation Program of China (2011DFA10440), and the Collaborative

Innovation Center of Resource-conserving & Environment-friendly Society and Ecological Civilization, the China Postdoctoral Science Foundation (No. 2014M552154), the Hunan Postdoctoral Scientific Program (No. 2014RS4026), and the Postdoctoral Science Foundation of Central South University (No. 126648).

## References

1. Li, S.Z.: Markov random field modeling in image analysis, p. 21. Springer-Verlag London Limited, UK (2009)
2. Nishino, K., Kratz, L., Lombardi, S.: Bayesian defogging. *International Journal of Computer Vision* 98(3), 263–278 (2012)
3. Kratz, L., Nishino, K.: Factorizing scene albedo and depth from a single foggy image. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1701–1708 (2009)
4. Fattal, R.: Single image dehazing. *ACM Transactions on Graphics* 27(3), 1–9 (2008)
5. Tan, R.T.: Visibility in bad weather from a single image. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8 (2008)
6. Carr, P., Hartley, R.: Improved single image dehazing using geometry. In: *The Digital Image Computing: Technique and Applications*, Melbourne, pp. 103–110 (2009)
7. Cataffa, L., Tarel, J.P.: Markov random field model for single image defogging. In: *IEEE Intelligent Vehicle Symposium*, pp. 994–999 (2013)
8. Cataffa, L., Tarel, J.-P.: Stereo reconstruction and contrast restoration in daytime fog. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) *ACCV 2012, Part IV*. LNCS, vol. 7727, pp. 13–25. Springer, Heidelberg (2013)
9. Wang, Y.K., Fan, C.T., Chang, C.W.: Accurate depth estimation for image defogging using Markov Random Field. In: *International Conference on Graphic and Image Processing (ICGIP)*, Singapore, pp. 1–5 (2012)
10. Kolmogorov, V., Zabin, R.: What energy function can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 26(2), 147–159 (2004)
11. Boykov, Y., Kolmogorov, V.: An experimental comparison of Min-cut/Max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 26(9), 1124–1137 (2004)
12. Rossum, M.V., Nieuwenhuizen, T.: Multiple scattering of classical waves: microscopy, mesoscopy and diffusion. *Reviews of Modern Physics* 71(1), 313–371 (1999)
13. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. *International Journal on Computer Vision* 48(3), 233–254 (2002)
14. The geo-v3.0 library (geo-v3.0), <http://vision.csd.uwo.ca/code/> (accessed April 5, 2013)
15. Boykov, Y., Veksler, O., Zabin, R.: Fast approximate energy minimization via graph cuts. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)* 23(11), 1222–1239 (2001)
16. Delong, A., Osokin, A., Isack, H.N., Boykov, Y.: Fast Approximate Energy Minimization with Label Costs. *International Journal of Computer Vision* 96(1), 1–27 (2012)
17. Fattal, R.: Single image dehazing, *ACM Transactions on Graphics (SIGGRAPH 2008)* 27, 72:1–72:9 (2008)
18. Tan, R.T.: Visibility in bad weather from a single image. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, United States, pp. 1–8 (2008)
19. He, K.M., Sun, J., Tang, X.O.: Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12), 2341–2353 (2011)

# An Improved Laparoscopic Image Registration Algorithm Based on Sift

Jiujiang Zhou, Jianxu Mao, and Xiaoyan He

College of Electrical and Information Engineering, Hunan University, Changsha, China  
jiujiang.z@gmail.com, mao\_jianxu@126.com

**Abstract.** Image registration is a recognized difficulty and many people are working on it to make their algorithms more efficient and robust. In image-guided surgical and interventional procedures, the registration precision and real time effect are both quite important for the following accurate tissue deformation recovery and 3D anatomical registration as well as navigation. This article uses the radon-transform and bidirectional matching approach on SIFT(Scale Invariant Feature Transform) which is aiming at the registration in laparoscopic binocular vision. Finally, we test the new algorithm and give better experiment results by comparing with other common methods.

**Keywords:** Medical Image Registration, Radon Transform, Bidirectional Matching, SIFT.

## 1 Introduction

Medical image registration is the basis to realize medical image information fusion and three-dimensional reconstruction and so on, it has been widely applied in disease diagnosis and preoperative planning, etc. Because its better adaptability to position change, gray level change, image distortion and complex space transform, the registration based on feature point is the current mainstream direction as well as development trend. A good feature detector or feature tracking technique is important for accurate tissue deformation recovery, 3D anatomical registration and navigation in computer assisted MIS(Minimally Invasive Surgical) procedures. Although there are a variety of methods which have been developed for image registration, there is no higher-performance and shorter time-consuming algorithm available for fitting great changes like scale, rotation, affine and projection.

Reference [1] expounds the review of medical image registration. Reference [2] raises Harris operator, which has a good performance with rotation and illumination changes but it is also variational with scale changes. Reference [3] gives the Scale Invariant Feature Transform(SIFT) algorithm and it is one of the best methods for feature detection and matching due to the invariance of scale, rotation, illumination, geometric distortion and resolution differences. Reference [4] introduces an Speeded Up Robust Features(SURF) which is based on integral image haar derivation and reduces the operation time. Reference [5] comes up with PCA-SIFT to reduce the dimension. However, this algorithm is not fully affine invariant and the projection matrix requires a series of representative

images. Recently, Harris and Hessian corner point detectors have also been extended to detect affine-invariant regions in [6] and [9], respectively. In [7], the salient region detector is proposed, where the local maximum in affine transformation space is detected by measuring the entropy of pixel intensity histograms computed for elliptical regions. Reference [8] uses a watershed like segmentation algorithm to detect Maximally Stable Extremal Regions (MSER) which is closed under the affine transformation of image coordinates and invariant to affine transformation of intensity. A comparison of these affine region detectors can be found in [9].

This article introduces SIFT in detail and make some improvements based on it. Section1 describes the image registration application on MIS and the commonly used image registration method based on feature points. Section2 gives the main process of SIFT algorithm. Section3 and Section4 demonstrate our modified ideas on two aspects: dimension reduction of the descriptor based on radon transform; a novel bidirectional matching approach. Section5 displays the experimental results and analysis, the method is applied on laparoscopic binocular vision images and the results show greater accuracy and faster matching. Section6 gives the conclusions and the deficiency of our method as well as the direction of further improvements. The purpose of this paper is to reduce the operating time as well as increase the efficiency of the algorithm for image registration in MIS and our proposed idea is tested on *in vivo* video sequences from robotic assisted MIS procedures.

## 2 Scale Invariant Feature Transform

The scale invariant feature transform (SIFT) algorithm, developed by Lowe[3,10,11], is an algorithm for image features matching which is invariant to image translation, scaling, rotation and partially invariant to illumination changes and affine projection. The main process of the algorithm is composed of the following four parts:

### 2.1 Scale-Space Local Extrema Detection

Lowe uses gaussian difference function to identify the key points which are scale and orientation invariant. The scale space of the image is defined as the function  $L(x, y, \sigma)$ , which is convolved by the variable scale gaussian function

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (1)$$

in which the gaussian function is:

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

Differential gaussian scale space  $D(x, y, \sigma)$  is defined as the convolution between the original image and the adjacent differential scale gaussian function which includes a constant factor  $K$ . These local extremum of gaussian difference image are regard as the feature points on the corresponding scale spatial domain, and the function  $D(x, y, \sigma)$  can be expressed by:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3)$$

## 2.2 Accurate Positioning of Feature Points

By calculating the fitting function

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \quad (4)$$

and further catching the extreme point

$$\hat{X} = -\left(\frac{\partial^2 D}{\partial X^2}\right)^{-1} \frac{\partial D}{\partial X}, \quad (5)$$

we can get the corresponding extreme value

$$D(\hat{X}) = D + \frac{1}{2} \frac{\partial D^T}{\partial X}, \quad (6)$$

thus obtain the local optimal point by revising  $X$  and get rid of the weak feature points in which

$$\left| D(\hat{X}) \right| < 0.03. \quad (7)$$

Meanwhile, accurate location and scale of the candidate feature points are acquired. The edge points also need to be removed by Hessian matrix. Assuming that  $\text{Tr}(H)$ ,  $\text{Det}(H)$  are adding result and product of the eigenvalues,  $\gamma$  is the ratio of these matrix eigenvalues, for the sake of testing if the main curvature is less than a certain threshold  $\gamma$ , just check if the equation

$$\frac{\text{Tr}(H)^2}{\text{Det}(H)} < \frac{(\gamma+1)^2}{\gamma} \quad (8)$$

is right, which  $\gamma$  is defined as 6~10.

## 2.3 The Generation of SIFT Feature Descriptor

Collect the gradient information and direction distribution property in the  $3\sigma$  neighborhood of the gaussian pyramid image, in which the radius is defined as:

$$\text{radius} = \frac{3\sigma_{oct} * \sqrt{2(d+1)+1}}{2}, \quad (9)$$

and confirm the gradient direction in the known keypoints. Ascertain the maximum value of histogram which is divided into 36 bins is the main direction of the keypoint and only keep the bins which have the size of the 80% of the peak value in the main direction or greater than it as the assistant direction of this keypoint which is to keep the descriptor's invariance of scale. The value and direction of the gradient at keypoint  $(x,y)$  are described as

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (10)$$

$$\theta(x, y) = \arctan \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} . \quad (11)$$

At last, the 128 dimensional vector is normalized for the invariance of the influence of illumination change, and after the processing the vector can be described as

$$L = (l_1, l_2, \dots, l_{128}) , \quad (12)$$

in which

$$l_i = \frac{h_i}{\sqrt{\sum_{j=1}^{128} h_j}} . \quad (13)$$

### 3 Descriptor Dimension Reduction

In this part, a novel formed descriptor is put forward based on radon transform [12]. The intrinsic quality of the registration based on keypoints between two images is that just matching two image subregions centered on every feature keypoint. As a consequence, first acquire the scale, orientation and position information of the feature points by using gaussian difference scale space. Second, applying the radon transform on the image to be processed in a series of straight line and catch the new descriptors for matching. By appropriately selecting some specific angles which are used for calculating radon transform value, the purpose of dimension reduction is achieved. At last, add feature point direction angle into the integral function to reduce the computational cost caused by rotation. The details are as follows.

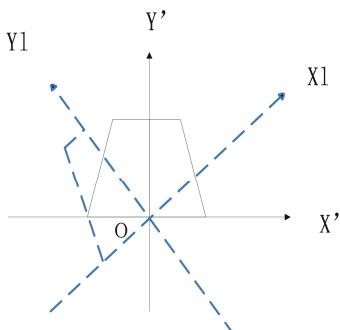
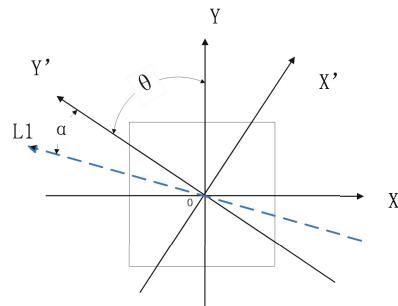
#### 3.1 Radon Transform Introduction

Assuming that the main direction of feature points detected by SIFT feature is  $Y'$ , while the angle is defined as  $\theta$  between  $Y'$  and y coordinate system. In original SIFT algorithm, in order to transform the tiny displacement of the feature points which results in the change of feature variable value, Lowe make a gaussian weighted smoothing method. When making a pixel gradient amplitude statistics, the sampling region near the feature point is highlighted. However, the proportion is dropping with the distance between the sampling region and the center point increasing. This paper will achieve the same purpose by changing product factor included by integral function of the image radon transform, and adopt the product factor  $\frac{1}{1+\sqrt{x}}$ . The integral

function is :

$$R(x) = \int_s \frac{1}{1+\sqrt{x}} I(x, y) \quad (14)$$

in which  $s'$  represents the integral region with the keypoint centered and is same to the statistic region in the original SIFT algorithm. The radon transform of the image is shown as Fig1.(a).

**Fig. 1. (a)** The radon transform**Fig. 1. (b)** The main direction of feature points and rotation Angle

### 3.2 Seek For D-Dimension Characteristic Vector Space

As shown in the Fig. 1(b), for the purpose of achieving d-SIFT feature vector space, feature point's principle orientation  $Y'$  is deemed to be as the reference direction. Make other  $d-1$  lines ( $L_1, L_2, \dots, L_{d-1}$ ) and thereinto,  $L_1$  is shown in the Fig. 1(b). The included angle between two adjacent lines

$$\alpha = \frac{2\pi}{d}, \quad (15)$$

so the radon transform of main direction line  $Y'$  belongs to image  $I(x, y)$  can be expressed by:

$$R_\theta(x) = \int_s \frac{1}{1 + \sqrt{x \cos \theta - y \sin \theta}} I(x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta) dy. \quad (16)$$

Similarly, the radon transform of  $I(x, y)$  along the lines  $L_1, L_2, \dots, L_{d-1}$  can be shown as:

$$R_{\omega_n}(x_{l_n}) = \int_s \frac{1}{1 + \sqrt{x_{l_n} \cos \omega_n - y_{l_n} \sin \omega_n}} I(x_{l_n} \cos \omega_n - y_{l_n} \sin \omega_n, x_{l_n} \sin \omega_n + y_{l_n} \cos \omega_n) dy_{l_n} \quad (17)$$

in which

$$\omega_n = \theta + n\alpha, n = 1, 2, \dots, d-1, \quad (18)$$

Fig. 1(b) shows the main direction of feature point of the image  $I(x, y)$  and radon transform of line  $L_1$ .

### 3.3 Selecting the Appropriate Integral Angle

In order to apply the idea into our method, choosing the integral angle  $\alpha$  as  $\frac{\pi}{15}$ , and normalized the length of eigenvector to unit form. Then, achieve 30 dimensional radon-SIFT descriptor:

$$R_\theta(x), R_{\omega_1}(x_{l_1}), R_{\omega_2}(x_{l_2}), \dots, R_{\omega_{29}}(x_{l_{29}}) \quad (19)$$

## 4 Bidirectional Matching Approach

### 4.1 Define the Distance of the Eigenvectors

Generally speaking, the best candidate match for a keypoint in  $I_1$  is found by measuring the nearest distance between keypoints. The minimum euclidean distance is defined for the invariant descriptor vector and used to match points from  $I_1$  to  $I_2$ . Supposing that the feature descriptor generated respectively from two images are:

$$R_\theta(x_1), R_{\omega_1}(x_{l_1}), R_{\omega_2}(x_{l_2}), \dots, R_{\omega_{29}}(x_{l_{29}}); R_\theta(x_2), R_{\omega_1}(x_{2l_1}), R_{\omega_2}(x_{2l_2}), \dots, R_{\omega_{29}}(x_{2l_{29}}), \quad (20)$$

thus euclidean distance will be:

$$d = \sqrt{\sum_{i=0}^{29} (R_{\omega_i}(x_{l_i}) - R_{\omega_i}(x_{2l_i}))^2}. \quad (21)$$

This paper still choose the Best Bin First(BBF) method to search the nearest and next nearest neighbor feature points, and the ratio between them is set a threshold value T. If the ratio is lower than T, the matching is conformed successfully.

### 4.2 Bidirectional Match

Suppose that F and G are two set of descriptors generated by  $I_1$  and  $I_2$  separately, and the size of them are M and N. Considering the feature similarity, We define the measurement function by  $S = \Omega(F, G)$  where S is the similarity measurement matrix, in which  $\Omega$  is method of similarity measurement. After obtaining S matrix, select a arbitrary element of F and make the comparison with all the elements in G, and then find out the pair which S obtain the extremum. Next, traverse in turn all the other elements of F as well as obtaining the correspondences, in which this process is a one-way mapping from F to G. The correspondence can be roughly shown as:

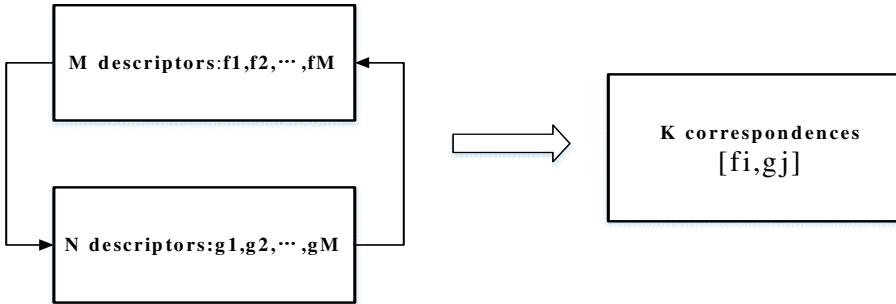
$$\text{correspondings}(S_{\max(1 \rightarrow N)}) = [\text{maximum}(F \rightarrow G)_{1 \rightarrow N}, \text{index}(F \rightarrow G)_{1 \rightarrow N}], \quad (22)$$

on the contrary, one-way mapping from G to F:

$$\text{correspondings}(S_{\max(1 \rightarrow M)}) = [\text{maximum}(G \rightarrow F)_{1 \rightarrow M}, \text{index}(G \rightarrow F)_{1 \rightarrow M}], \quad (23)$$

where the maximum  $(F \rightarrow G)_{1 \rightarrow N}$  represents the max value of S, index  $(F \rightarrow G)_{1 \rightarrow N}$  represents the corresponding descriptors, and vice versa. Finally, gain the best K matches which is approximately shown as figure2 form the two descriptor sets. The K should also meet the Eq.(24)

$$[index(F \rightarrow G)_{1 \rightarrow K}] \& [index(G \rightarrow K)_{1 \rightarrow K}] = K \quad (24)$$



**Fig. 2.** Two-way matching diagram

#### 4.3 A Mixed Similarity Measurement Strategy

The previous section discusses the matching method and the way of similarity measurement  $\Omega$  is used. In the field of medical image registration, the evaluation of registration results does not exist the so-called absolute gold standard, which is due to illumination change, imaging equipment parameters, etc. With the continuous development of research on image registration there have been some good similarity measure functions, such as quadratic sum of the gray differentials, mutual information, image correlation and so on. Due to the application for registration in the laparoscopic binocular vision images are mostly center symmetrical, the global measurement function  $\Omega$  is determined by combining image gray level information and feature descriptors at the same time. Therefor, we also set up a weight value  $\beta$  for optimum proportion. The measurement function is:

$$\Omega[I_1, I_2] = \beta\Omega_\beta(p(x), q(y)) + (1-\beta)\Omega_{1-\beta}(f(x), g(y)). \quad (25)$$

Among these symbols,

$$\begin{cases} \Omega_\beta(p(x), q(y)) = \frac{1 - |p(x) - q(y)|}{R} \\ \Omega_{1-\beta}(f(x), g(y)) = \frac{\langle f(x) \bullet g(y) \rangle}{\|f(x)\| \|g(y)\|} \end{cases}, \quad (26)$$

$p(x)$  and  $q(y)$  are the gray information of image  $I_1$  and  $I_2$ ,  $f(x)$  and  $g(y)$  are feature descriptors of these two images respectively, R is the gray statistics region.

$\langle \bullet \rangle$  is the distance between the descriptors as described above,  $\beta$  is adjustment coefficient.  $\Omega_\beta(p(x), q(y))$  is gray similarity measure function,  $\Omega_{1-\beta}(f(x), g(y))$  is descriptor similarity measure function. To continuously verify registration results parameter  $\beta$  needs to be adjusted. In this paper we determine that the value is 0.25 based on the experimental results of specific laparoscopic environment.

## 5 Experimental Results

The sensitivity of a given registration algorithm is estimated as defined sensitivity as the ratio of correct matchings and all the detected matchings.

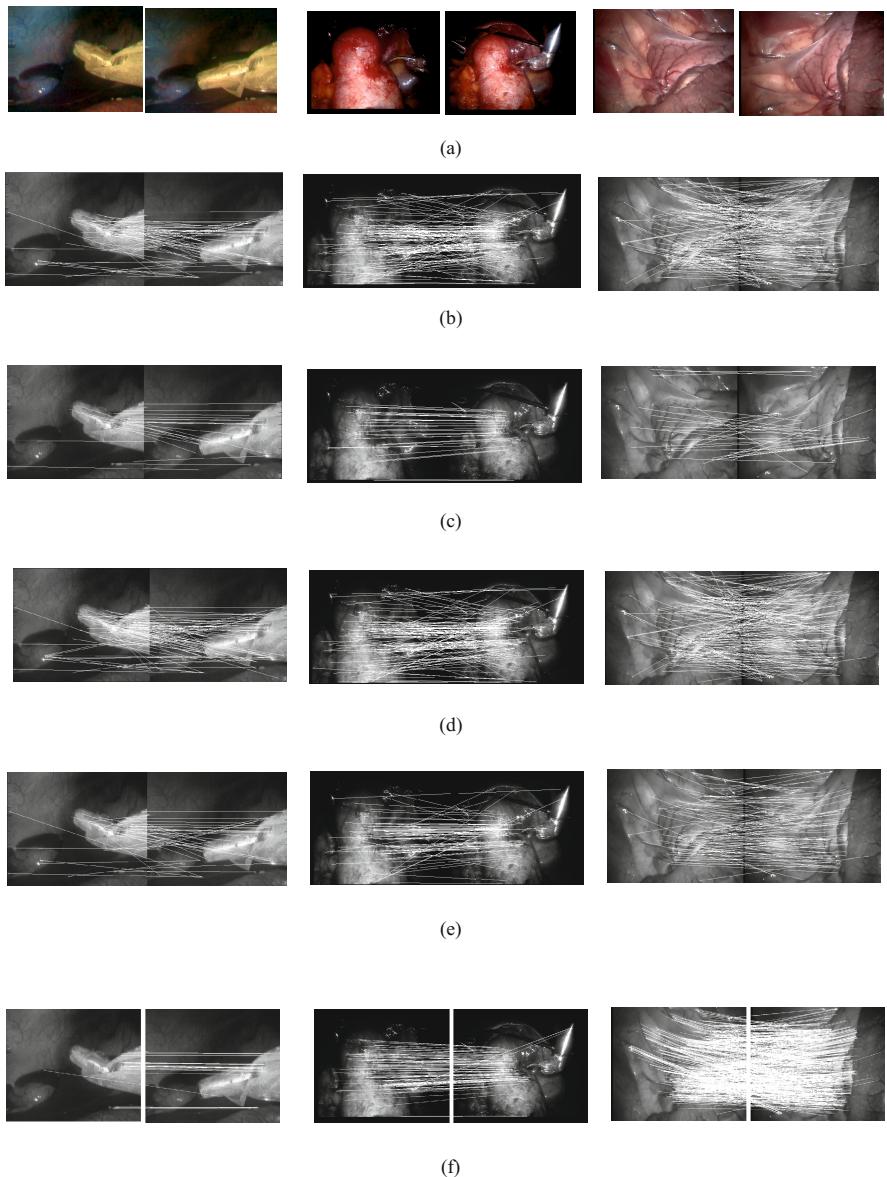
$$\text{sensitivity} = \frac{\# \text{correct\_matches}}{\# \text{correspondences}} , \quad (27)$$

In case of illumination, scale and rotation changes, test the new improved algorithm on the given data [13], as is shown in Fig3 (a), which is aiming at the improvement of characteristic dimension and matching strategy. At last, the behavior of new proposed algorithm is compared with several other classical algorithms. Experimental platform: CPU: 2.7 GHz core i5, memory 6G, Windows 7, Matlab2010. Through trial and error, the datum obtained are calculated as follows:

**Table 1.** The comparison diagram of different algorithm

	SIFT		SURF		PCA-SIFT		MSER		OUR METHOD	
	Time (s)	Sensitivity (%)	Time (s)	Sensitivity (%)	Time (s)	Sensitivity (%)	Time (s)	Sensitivity (%)	Time (s)	Sensitivity (%)
Illumination change	4.81	91.5	3.15	83.1	2.67	87.2	3.62	68.8	2.15	92.8
Scale change	3.38	84.6	1.98	87.7	2.19	82.8	2.97	76.6	1.86	89.7
Orientation change	6.07	73.1	3.12	65.8	3.8	66.9	4.5	84.6	3.25	81.5

The Table1 shows that under the three common transformations, compared with the original sift algorithm and other commonly used algorithm, the new 30-d radon bidirectional-SIFT has better overall promotion at the accuracy of the matching as well as time reduced by about 30%. The second way SURF depends on local pixel gradient direction much, and it may also be invalid as long as the layers of the pyramid are not closely enough. Although PCA-SIFT reduces the running time, but the projection matrix still needs some typical images. For MSER, its main advantage is good affine invariance. At the same time, this algorithm needs more improvement on instantaneity. In general the new algorithm improves the efficiency especially on registration accuracy under conditions of keeping good scale invariance and the original sift matching ratio. The results of the experiment effect using our novel algorithm can be visible as the following figures:



**Fig. 3.** Test images and results. (a) Original Images. (b) SIFT Results. (c) SURF Results. (d) PCA-SIFT Results. (e) MSER Results. (f) Our Improved Algorithm Results.

## 6 Conclusions and Discussion

In this paper, a novel feature detection and matching algorithm based on SIFT is presented. Experimental results have shown that the proposed improvement of the

original method is capable of making a more accurate matching and taking less time. This work improves the SIFT contrapositing the matching accuracy as well as operating time and provides some new thinking for improving the efficiency of algorithm. The SIFT itself maybe also have limitations, maybe it relies too much on the gradient direction of local pixels when striving for main direction which results in matching errors increasing; just make up the scale error when building dimension by interpolation. In the following work, the emphasis will continue to be put on searching for more efficiency algorithm and prepared for the higher precision registration requirement in 3d reconstruction. Of course it can also be used for *in vivo* tracking, significant deformation and intraoperative registration.

**Acknowledgements.** This work is supported by National Natural Science Foundation of China (61072121, 60835004, 61271382), Hunan Provincial Natural Science Foundation of China (12JJ2035) and the Fundamental Research Funds for the Central Universities, Hunan university.

## References

- [1] Maintz, J.B., Viergever, M.A.: A survey of medical image registration. *Medical Image Analysis* 2(1), 1–36 (1998)
- [2] Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, p. 50 (1988)
- [3] Lowe, D.G.: Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
- [4] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
- [5] Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II-506–II-513. IEEE (2004)
- [6] Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. *International Journal of Computer Vision* 60(1), 63–86 (2004)
- [7] Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Pajdla, T., Matas, J(G.) (eds.) *ECCV 2004*. LNCS, vol. 3021, pp. 228–241. Springer, Heidelberg (2004)
- [8] Matas, J., Chum, O., Urban, M., et al.: Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing* 22(10), 761–767 (2004)
- [9] Mikolajczyk, K., Tuytelaars, T., Schmid, C., et al.: A comparison of affine region detectors. *International Journal of Computer Vision* 65(1-2), 43–72 (2005)
- [10] Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
- [11] Lowe, D.G.: Local feature view clustering for 3D object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, pp. 682–688 (December 2001)
- [12] Kadyrov, A., Petrou, M.: The trace transform and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(8), 811–828 (2001)
- [13] Giannarou, S., Visentini-Scarzanella, M., Yang, G.-Z.: Probabilistic Tracking of Affine-Invariant Anisotropic Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 99 (2012)

# Application of Image Processing Techniques in Infrared Detection of Faulty Insulators

Yefan Wu<sup>1,\*</sup>, Jiangang Yao<sup>1</sup>, Tangbing Li<sup>2</sup>, Peng Fu<sup>1</sup>, Wei Liao<sup>1</sup>, and Mi Zhang<sup>1</sup>

<sup>1</sup> College of Electrical and Information Engineering, Hunan University, Changsha, China  
wuyefanbie@gmail.com

<sup>2</sup> Jiangxi Electric Power Research Institute, Nanchang, China

**Abstract.** The image processing technique is an essential way to ensure an accurate infrared detection of faulty insulators. In this paper, we analyze the necessity of images processing techniques in infrared detection of faulty insulators, research the related image processing techniques applied in infrared detection of faulty insulators and provide the corresponding practical examples. The work have done in this paper can make a contribution to the application of image processing techniques in infrared detection of faulty insulators.

**Keywords:** image processing, infrared thermal image, insulators.

## 1 Introduction

Insulator detection is an onerous work of the current power line patrol [1], which would take a plenty of manpower, material resources and time [2]. Low and zero value insulators detection mainly depend on artificial lever operation, having a certain risk and being highly influenced by the environment. Aiming to reduce the risk and intensity of the work, a plenty of researches have been done by scholars overall worldwide. A simple, accurate and economic on-line detecting method of insulators is need at present.

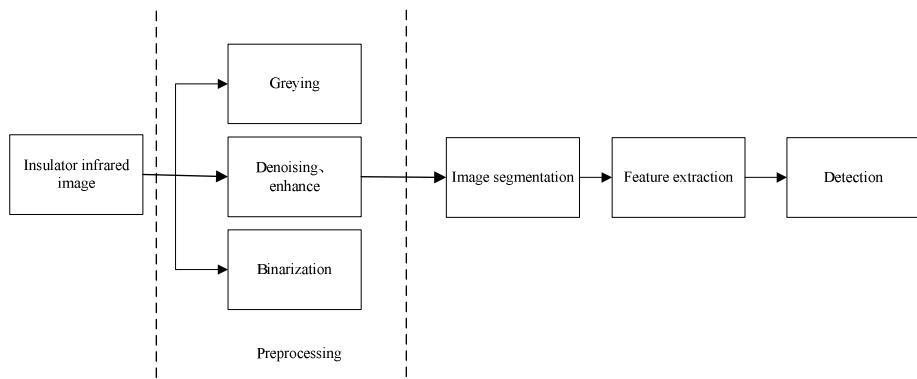
With the development of modern infrared technique, the infrared detecting technique of electrical equipment condition utilizing the advantages of infrared detecting such as accuracy, real-time and celerity etc., is developing at a great rate [3,4,5]. Compared with the traditional detecting technology, it has more advantages such as non-contact, high safety and convenient operation etc., having been widely used in faulty detection of electrical equipment [6,7,8]. However, the infrared thermal images still have some disadvantages such as low contrast between the object and background, edge blur and the noise [9]. The clear infrared thermal images and intact image details are the key to make an accurate diagnose of zero value insulators. Thus, to ensure an accurate detection of faulty insulators by infrared detection we need to do some effective processing of insulator infrared images.

This paper is organized as follows. The necessity of images processing techniques in infrared detection of faulty insulators is analyzed in Section 2. In Section 3, the related image processing techniques applied in infrared detection of faulty insulators

are presented, and the corresponding practical examples are also provided to testify the feasibility of the image processing techniques. Some concluding remarks are presented in Section 4 of the paper.

## 2 The Necessity of Image Processing Techniques in Infrared Detection of Faulty Insulators

There is a strict requirement of image processing technology in infrared detection of faulty insulators. Because the effect of the image processing directly affects the final test results. However, the infrared images of the scene can not be used directly for judgment, which should be processed by some techniques. The corresponding process of image processing is shown in Fig. 1.



**Fig. 1.** The process of image processing techniques in infrared detection of faulty insulators

### 2.1 The Necessity of Preprocessing

The clear view, complete image details and characteristics of insulator in infrared thermal image are very important to the faulty insulators detection. In the process of image generation, infrared thermal image is influenced by all kinds of electronic devices and detector noise, having the characteristics of high noise and low contrast. Part of the image details and characteristics masked by noise, causing the infrared image degradation, directly affect the accuracy of the follow-up work such as the insulator infrared image segmentation, feature extraction, fault judgment. The preprocessing is shown in Fig. 1.

### 2.2 The Necessity of Image Segmentation

In the research of the faulty insulator infrared thermal image, the infrared image not only has the insulator, but also includes other distractors such as background, wires, experimental simulation conductor. However, the research object is insulator disk or

steel cap area. If we don't do the infrared image segmentation, we would extract some error pixels, which affect the accuracy of faulty insulator inspection and positioning. Therefore, the image segmentation is an important step in the detection of faulty insulators based on the infrared thermal image technology.

### 3 The Application of Image Processing in the Infrared Detection Method

#### 3.1 Preprocessing

Infrared thermal image noise mainly comes from the infrared focal plane array of sensors, electronic circuit noise, background noise, etc. Therefore, we can analysis of the infrared image noise through the analysis of the noise of infrared focal plane imaging system. There usually have two kinds of infrared focal plane array of noise: the inherent spatial noise and the transient noise. In the earlier references, generally they believe infrared focal plane array deriving from the typical noise mechanism, including heat detector noise, shot noise, noise, etc.[10]. In the insulator infrared thermal image, the difference of temperature between the area of insulator and the background is small, and the dynamic range is large. In order to accurately identify the faulty insulator, enhanced processing of infrared thermal image is needed. Histogram equalization is a common image enhancement method. Transformation function is used to transform the original image to a uniform probability distribution. It can enhance pixels which have large proportion and inhibit pixels which have little proportion [11]. Platform histogram equalization method is an improve method of histogram equalization method, selecting an appropriate platform threshold, if a grayscale platform for the probability distribution is greater than the threshold, we put the probability distribution for the threshold; if the probability distribution is less than the threshold platform, we remain the original threshold [12].

Several common principles of denoising method used in infrared detection technology of faulty insulator can be expressed as follows:

##### 3.1.1 The Median Filter

Median filtering is a nonlinear processing technology, as well as the most common image processing technique used in preprocessing technology [13,14,15]. Median filter can overcome fuzzy linear filter and effectively filter out pulse the interference and the noise of image scanning. It also can get rid of the noise in the image edge. But for some image with many details, especially dot, line, peaked, median filter is unfavorable because it would lose some edge details.

The main role of median filter is to replace the pixels which have big difference between surrounding pixel gray value, by the pixels closed to surroundings. The property of denoising is also remarkable, especially for random noise and impulse noise.

### 3.1.2 Wavelet Transforms Denoising

Wavelet transform possessing a plenty of advantages such as time-frequency characteristic, low entropy, multi-resolution, the decorrelation and the flexibility of selected base have been widely used in image denoising.

The wavelet denoising method can be roughly classified into three categories: wavelet shrinkage method, projection method and all kinds of relevant improving methods. It is mainly used in two aspects: firstly, it achieves the goal of image denoising by adjusting the wavelet coefficients. Secondly, it improves the performance of the original image denoising algorithm by using time-frequency localization and multi-resolution of wavelet transform.

### 3.1.3 The Wiener Filtering Method

Wiener filtering method used the relevant characteristics of stationary random process and frequency spectrum to filter the signal mixed with noise. It regards minimum mean square error as the performance standards; the ultimate goal is to make the square error between processed image and original image minimal. Filtering effect of this method is better than the mean filter, it is useful for retaining the image edge and other high frequency part, but has large amount of calculation. Wiener filter has the best filtering effect with white noise image.

### 3.1.4 Neighborhood Average Method

Neighborhood average method is a typical linear filtering algorithm, and its basic principle is to use of the individual pixel values of average instead the original image. Dealing with the current pixel  $(x, y)$ , chooses a template which is composed of the neighbor number of pixels, and gives the average of all pixels in the template to the current pixel  $(x, y)$  as the image grey value  $g(x, y)$  on that point after processing.

$$g(x, y) = \frac{1}{M} \sum_{(x, y) \in s} f(x, y) \quad (1)$$

Where  $s$  is the template region of current pixel,  $M$  is the total number of the current pixel this template contains.

Using average filtering of neighborhood average method is very suitable for removing particulate noise in the image obtaining by scanning. Neighborhood average method effectively suppresses the noise, but also caused the fuzzy phenomenon due to the average fuzzy degree is proportional to the area radius.

### 3.1.5 Practical Example

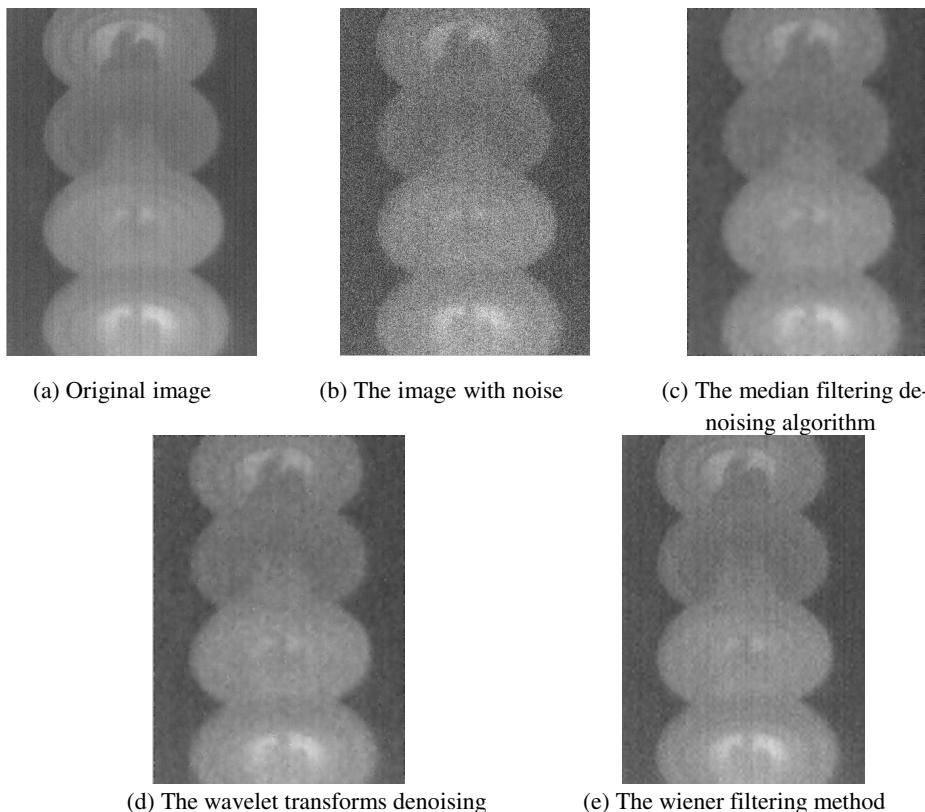
In order to verify the validity of the denoising methods above. In this paper, adds the gaussian white noise with mean square error of 5,10,15,20,25 respectively to insulator infrared images, and uses the different noise reduction method. The Table.1. shows the PSNR of different noise reduction method.

Usually, the bigger PSNR value is the better denoising effect, from The table 1 shows the Wiener Filtering Method calculated the PSNR values are higher than other methods in various noise level.

**Table 1.** The PSNR of different noise reduction method

	$\sigma=5$	$\sigma=10$	$\sigma=15$	$\sigma=20$	$\sigma=25$
The Median Filter	18.647	16.154	14.458	13.473	12.679
Wavelet Transforms Denoising	19.514	17.087	15.318	14.271	13.265
Neighborhood Average Method	19.926	18.163	16.214	15.131	13.891
The Wiener Filtering Method	20.623	19.021	17.226	16.010	14.718

But the objective indicators and human visual observation is not always consistent, therefore, we take a picture of insulators added white noise with mean square error of 15, and use the above denoising method deal with the picture, the visual effect picture of denoising results are shown in Fig. 2.

**Fig. 2.** The images before and after denoising

The results show that the insulator infrared image processed with an effective filtering method can denoise effectively, which is favorable for the later image segmentation and fault identification.

### 3.2 Segmentation

There are many different kinds of image segmentation methods put forward by some researchers for different fields and engineering applications. Generally, image segmentation methods can be divided into three categories, segmentation method based on threshold, region and the edge, respectively.

#### 3.2.1 Image Segmentation Method Based on Threshold

Threshold segmentation utilizes the difference of gray scale between the target and the background to regard the image as the combination of two different grayscale images. Selecting a suitable gray threshold and determining the each pixel point of image whether belonging to the target or background region, to generate the corresponding binary image. According to 3.1, the image  $g(x, y)$  can be separated as follows:

$$g(x, y) = \begin{cases} 1 & f(x, y) \geq t \\ 0 & f(x, y) < t \end{cases} \quad (2)$$

Where  $f(x, y)$  is the original image;  $t$  is the grey value orientation threshold.

Obviously, the selection of threshold is the key to the threshold segmentation. If threshold is too big, a lot of goals will be mistaken for background; if threshold value is too small, we will lose the target area. Several image segmentation methods based on threshold would be introduced as follows.

##### (1) The Otsu Method

Otsu method is a kind of adaptive segmentation method based on threshold. The optimal threshold is determined by the maximum interclass variance of gray histogram.

The probability of the pixel in grayscale  $i$  can be expressed as follows:

$$p_i = n_i / N \quad (i = 1, 2, \dots, L-1) \quad (3)$$

$$\sum_{i=0}^{L-1} p_i = 1 \quad (4)$$

Where  $n$  is the pixel of the image;  $L$  is the number of grey scales.

The grayscale image is divided into the background  $B$  and target  $O$  by the threshold  $t$ . The probability  $\omega_0$  and the average gray level  $\mu_0$  of the background part  $B$  expressed in equation (5); the probability  $\omega_1$  and the average gray level  $\mu_1$  of the background part  $O_1$  expressed in equation (6).

$$\omega_0 = \sum_{i=0}^t p_i, \quad \mu_0 = \sum_{i=0}^t ip_i / \omega_0 \quad (5)$$

$$\omega_1 = \sum_{i=t+1}^{L-1} p_i, \quad \mu_1 = \sum_{i=t+1}^{L-1} ip_i / \omega_1 \quad (6)$$

Therefore,  $\mu$  can be expressed as follows:

$$\mu = \omega_0 \mu_0 + \omega_1 \mu_1 \quad (7)$$

Interclass variance can be defined as follows:

$$\sigma^2 = \omega_0 (\mu_0 - \mu)^2 + \omega_1 (\mu_1 - \mu)^2 = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \quad (8)$$

The optimal threshold is the one which can make  $\sigma^2$  maximum.

### (2) The Iterative Threshold Method

Iterative threshold method selects a threshold for the initial value firstly, and then improves the estimation by some strategies, until meeting the set standards [16]. Median iteration method is one of the representative iterative method, the steps are as follows:

- 1) Choosing the image grey value as the initial threshold  $T_0$ ;
- 2) Dividing the image into two areas  $R_1$  and  $R_2$  through the threshold  $T$ , and calculating the shade of gray average  $\mu_1$  and  $\mu_2$  of the area  $R_1$  and  $R_2$  respectively as equation (9).

$$\mu_1 = \frac{\sum_{i=0}^{T_0} i n_i}{\sum_{i=0}^{T_0} n_i}, \mu_2 = \frac{\sum_{i=T_0}^{L-1} i n_i}{\sum_{i=T_0}^{L-1} n_i} \quad (9)$$

- 3) Calculating the new threshold  $T_{i+1}$ :

$$T_{i+1} = \frac{1}{2}(\mu_1 + \mu_2) \quad (10)$$

- 4) Repeating steps 2 and 3 until the difference between the threshold  $T_{i+1}$  and  $T_i$  is smaller than a certain value.

The iterative threshold method is simple, but still under the influence of threshold selection strategy. How to choose an appropriate threshold still needs further study.

### (3) The Maximum Entropy Threshold.

The maximum entropy utilizing entropy makes quantitative measure in the original image to ensure the entropy of target and background as large as possible after segmentation. It is generally believed that after the segmentation of image, the greater entropy value the more information we can get from the original image to ensure good overall segmentation effect [17], entropy is defined as:

$$H(A) = \sum_{i=1}^N p_i \log(p_i) \quad (11)$$

Where  $p_i$  is a probability density function random of variable  $i$  (grey value, regional gray or gradient).

### 3.2.2 Image Segmentation Method Based on Edge

Image edge is composed of discontinuous local image's properties, such as the mutation of color, grayscale and texture. Image segmentation method based on edge is

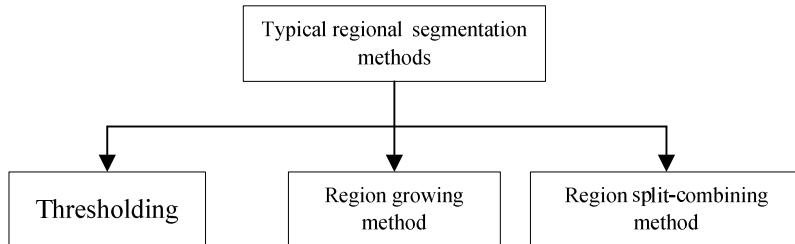
realized by detecting the edges of different regions. The main methods of image segmentation and their advantages as well as the disadvantages are presented in Table 2.

**Table 2.** The advantages and disadvantages of different image segmentation methods based on edge

Methods	Advantages	Disadvantages
Sobel operator method	The high accuracy of edge positioning	It is easy to cause the appearance of multi pixel width at the detection edge
Prewitt operator method	The high accuracy of edge positioning	It is easy to cause the appearance of multi pixel width at the detection edge
Roberts operator method	The high accuracy of edge positioning	It is easy to cause the loss of part of the edge and does not have the capability of suppression noise
Hough transform method	The operation is easy and have the capability of suppression noise	It takes a lot of storage space and computing time
Edge tracing method	The operation is easy	The improper search direction will cause errors of edge tracking

### 3.2.3 Image Segmentation Method Based on Region

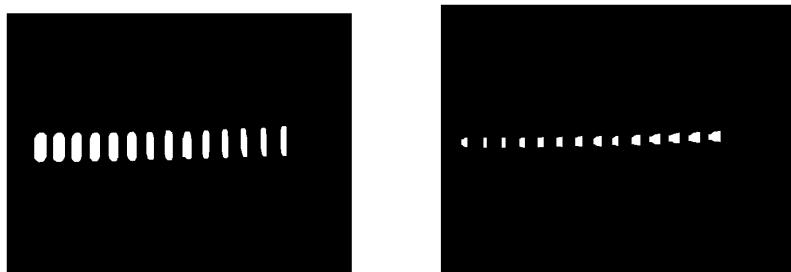
The essence of the regional segmentation is to connect the pixels which have the similar properties, and then forms objective segmented regions. The typical regional segmentation methods are shown in Fig. 3.



**Fig. 3.** The typical regional segmentation methods

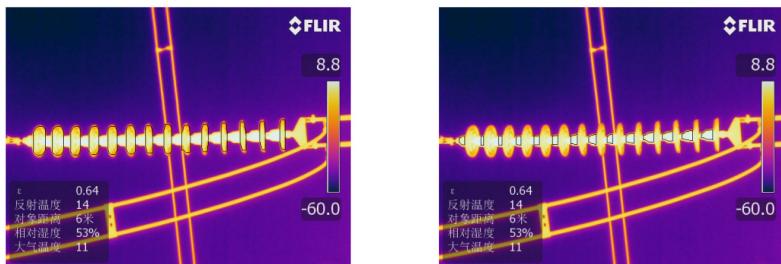
### 3.2.4 Practical Example

In order to test the effect of image segmentation, a practical example is employed. Fig. a and Fig. b in Fig. 3 are the binary images of disks and steel caps, respectively. Fig. a and Fig. b in Fig. 4 are the segmentation images of disks and steel caps, respectively.



(a) The binary image of disks

(b) The binary image of steel caps

**Fig. 4.** The binary images

(a) The segmentation image of disks

(b) The segmentation image of steel caps

**Fig. 5.** The segmentation images based on regional segmentation methods

From the figures above, we can see the effect of the image segmentation is favorable. We can get a clear area of disks and steel caps after the image segmentation, and then we can get the right information out of the infrared picture to do the further detection work of faulty insulators.

## 4 Conclusion

According the analysis and the practical examples presented in this paper, it is obviously that image processing technology can make a contribution to infrared detection of faulty insulators. This paper introduced the image preprocessing and segmentation of faulty insulators images which is favorable to the application of image processing techniques in infrared detection of fault insulators.

**Acknowledgement.** The authors are grateful for the support provided by the Foundation of Science and Technology Project of Jiangxi Power Corporation (Grant No. 201350617) and we are also thanks for the help of the organization and individuals whose literatures have been cited in this paper.

## References

1. Min, D.: Reliability appraisal of insulators for extra high voltage transmission line. *Power System Technology* 23, 40–41 (1995)
2. Yi, H.: Operating Current Situation and Prospect of Insulator for Transmission Line in China. *Electrical Equipment* 6, 1–4 (2005)
3. Chen, Y., Cai, K., Liu, Y.: The infrared diagnosing technology of power supply equipment, pp. 6–9. China Water Power Press (2006)
4. Zhang, Y., Yu, F.: Brief Introduction of Examining Procelain Insulator's Operation Status by Using the Infrared Thermal Image Technology. *Qinghai Electric Power* 22, 40–43 (2003)
5. Chen, H., Chen, Y., Zhao, X., et al.: Application of Infrared Temperature Measurement Technology in Detection of Composite Insulator. *Electrical Equipment* 7, 42–43 (2006)
6. Zhang, Q., Liu, H., Huang, X., et al.: An expert system for infrared fault diagnosis of power transformer. *Power System Technology* 26, 18–21 (2002)
7. Zhu, J., Wang, Y., Cui, S., et al.: The Application of Infrared Diagnosis Technology in Diagnosis of the High Voltage Electrical Equipment Internal Defect. *High Voltage Engineering* 30, 34–36 (2004)
8. Hu, S., Shen, X.: An infrared diagnostics example of internal moistened arrester. *Power System Technology* 20, 43–44 (1996)
9. Tian, Y.: Infrared detection and diagnosis technology, pp. 144–146. Chemical Industry Press (2006)
10. Xu, N., Bian, N.: The infrared radiation and guidance, pp. 206–211. National Defense Industry Press (1997)
11. Liu, Z., Li, Z.: A Review on Image Process Technique of Thermal Imager. *Infrared Technology* 2, 27–32 (2000)
12. Song, Y., Shao, X., Xu, J.: New enhancement algorithm for infrared image based on double plateaus histogram. *Infrared and Laser Engineering* 37, 308–311 (2008)
13. Guan, X., Zhao, L., Tang, Y.: Mixed Filter for Image Denoising. *Journal of Image and Graphics* 10, 332–337 (2005)
14. Zhang, X., Xu, B., Dong, S.: Adaptive switching median filter for the removal of impulse noise. *Opto-Electronic Engineering* 33, 78–83 (2005)
15. Jin, L., Xiong, C., Li, D.: Adaptive center-weighted median filter. *Journal of Huazhong University of Science and Technology (Nature Science Edition)* 36, 9–12 (2008)
16. Yao, M.: Digital image processing, pp. 225–262. China Machine Press (2006)
17. Tang, X., Zhang, X., Zou, H.: Improvement of the maximum entropy image segmentation method. *Computer Engineering and Applications* 48, 216–219 (2012)
18. Xiong, Z., Xiao, G., Qiu, K.: Watershed algorithm based on adaptive marker extraction and energy equation. *Computer Engineering and Applications* 48, 186–189 (2012)
19. Yu, S., Zhou, Y., Zhang, R.: Digital image processing, pp. 307–309. Profile of Shanghai Jiao Tong University Press (2007)

# Finding the Accurate Natural Contour of Non-rigid Objects in Video\*

Gaoxuan Ying, Sheng Liu, and Yiting Jin

College of Computer Science and Technology, Zhejiang University of Technology  
Hangzhou, 310023, Zhejiang, P.R. China  
`{edliu@zjut.edu.cn}`

**Abstract.** Non-rigid object tracking is an important task in computer vision, while its natural contour extraction is one of the most difficult problems during the process. Most tracking-by-detection methods are based on rectangular bounding-boxes, this will lead errors into subsequent detection. This paper present a novel superpixel-based detector for accurate natural contour extraction, there are three main contributions: 1) combining real-time superpixel segmentation with natural contour detection, 2) proposing an object-oriented natural contour extraction method for non-rigid objects, 3) proposing a non-rigid object detection method based on flexible scanning window. Compared with those bounding-box based detection methods, our detector can provide very accurate initial input of object model, then produce accurate natural contour output of the non-rigid object. Our detector broke the conventional detection method based on scanning rectangle, which greatly reduced the interference caused by background information. The experiments show that the proposed method outperforms the state-of-the-art algorithms not only on the contour accuracy but also on the computation cost. In addition, the initialization stage of our method overcomes the limitation of HT caused by the size of initial bounding-box.

**Keywords:** superpixel, non-rigid object, natural contour extraction, flexible scanning window.

## 1 Introduction

Object tracking is a major component in computer vision. It plays quite an important role in intelligent transportation systems, intelligent surveillance systems and military object detection, etc[1]. Certainly, object detection is quite a challenging research, because the appearance of the same target will change differently under different illumination or viewpoints. Especially for non-rigid targets, the object itself will deform with the flow of the video sequence, which will bring great difficulties to non-rigid object detection accompanied by the change of surroundings.

---

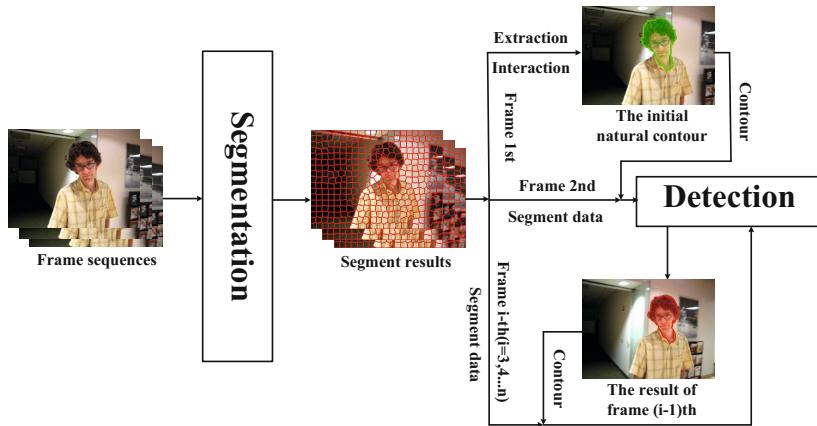
\* This work was supported by the National Natural Science Foundation of China (NSFC-60573123,60605013,60870002, 60802087), NCET, and the Science and Technology Dept of Zhejiang Province (2012R10052,Y1110688).



**Fig. 1.** Some results of our detector (green: initialization; red: result)

So far, non-rigid object detection and tracking methods can be divided into two categories by the form of the tracking results. One is the method with a bounding-box as output[2][3][4][5][6][8][9], the other is the method with a natural contour as output[7]. The former is limited to a bounding box representation with fixed aspect ratio, or modeling the non-rigid object with a couple of bounding-boxes. Thus, it lacks precision in segmenting the object from background because the detection results contain much background information, and this will lead errors into subsequent detection, therefore, the tracking results will be inadequate to perform tasks that demand a high degree of accuracy, such as object recognition and behavior analysis. The latter is based on a natural contour, this kind of method can avoid the problems brought by the bounding-box. The detection results can segment out the target from background with a contour that closely follows the object boundary, it can provide a reliable calculation basis for some high-level computer vision tasks, such as recognition, animation, behavior analysis, etc. Hence, it can greatly improve the performance of the surveillance systems.

In the process of detection based on natural contour, the key problem is how to segment out the non-rigid object from background. To solve this problem, Martin Godec[7] proposed a non-rigid object tracking algorithm based on the Hough-transform (HT), they extend the idea of hough forests to the online domain and couple the voting based detection and back-projection with a rough segmentation based on GrabCut. This significantly reduces the amount of noisy training samples during online learning and thus effectively prevents the tracker from drifting. Although, the algorithm can avoid the interference of background information brought by rectangular box, it will affect the accuracy of the segmentation results directly if node-voting errors occur. Just like HT, most non-rigid object tracking algorithms are on the pixel level[2][3][9], this will make the time complexity of the algorithm very high.



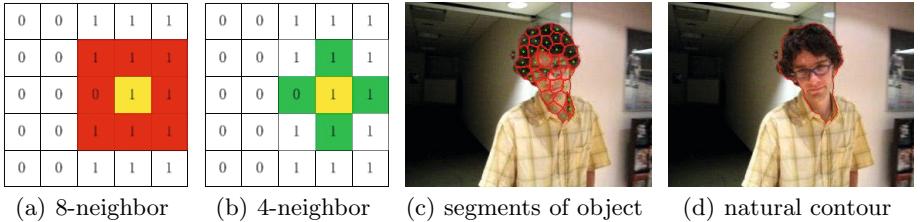
**Fig. 2.** Overview of our detector

In this work, we proposed a superpixel-based detector for non-rigid objects. The proposed algorithm has two advantages. Firstly, we use superpixels instead of pixels which results in a sharp decline in the amount of processing data. Secondly, the natural contour of non-rigid object is consists of some boundary points of superpixels, therefore, we can catch accurate natural contour of the non-rigid object and achieve accurate natural contour detection.

## 2 Overview

We propose a natural contour detection method based on real-time superpixel segmentation. Firstly, we propose an object-oriented natural contour extraction method to catch the natural contour of non-rigid objects. Then, we propose a detection method based on flexible scanning window, thus realize the precise tracking of non-rigid object. The framework of proposed algorithm is shown in figure 2. As can be seen in the framework, we use a real-time superpixel segmentation method to preprocess the video frames, then with a simply human interaction, catch the accurate initial natural contour of the non-rigid object by our proposed NCE (Natural Contour Extraction) method, this part will be introduced in section 2.1. To perform the scanning and detecting tasks well, we take both the initial natural contour and the segment data of current frame as the inputs of our detection algorithm. Then, the output contour will be detected based on flexible scanning window, this part will be introduced in section 2.2. There are three main contributions in this paper,

1) We combined real-time superpixel segmentation with natural contour detection. Using superpixels instead of pixels can reduce the calculation , and superpixels provide very favorable boundary structural information for natural contour detection.



**Fig. 3.** Natural Contour Extraction algorithm: (a) is the rule in SLIC, (b) is the rule of NCE, (c) is the input of NCE and (d) is the output

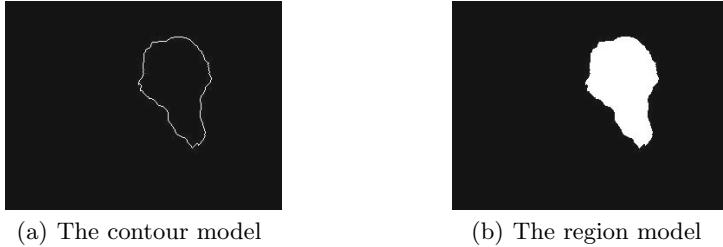
2) We proposed an object-oriented natural contour extraction method of non-rigid objects. We extract and combine the boundary information of segments those belong to the initial object by NCE to get the initial natural contour of non-rigid objects.

3) We proposed a non-rigid object detection method based on flexible scanning window. Using irregular contour curve to scan and detect, we broke the conventional detection method based on scanning rectangle, which greatly reduced the interference caused by background information.

## 2.1 Natural Contour Extraction

As a kind of preprocessing method, Superpixel segmentation is quite important to missions of computer vision because it can extract features of mid-level. But if it is needed to be applied to practice, the segmentation algorithm should result in high quality (low under-segmentation error and high boundary recall) superpixels with control on its number, and also have low computational cost. Through the study of some superpixel segmentation algorithms, we found that SLIC (simple linear iterative clustering)[10] and Turbopixels[11] these two segmentation algorithms are suitable for our detector. They both are  $O(N)$  complex and can control the number of superpixels. And the boundary computed by them is quite fit with the truly boundary. We test both of them with the same dataset, the result shows SLIC has higher boundary recall and lower under-segmentation error which means the quality of SLIC superpixels is higher than Turbopixels. Besides, we can use GPU to realize a real-time SLIC superpixel segmentation (gSLIC)[12]. In conclusion, we chose SLIC segmentation algorithm as our preprocessing method. With plenty of testing, we found different suitable SLIC parameters which can lead to high quality segments of videos with different frame size.

After preprocessing the frame by SLIC segmentation, every point got a label that shows which segment it belongs to. We proposed an algorithm called Natural Contour Extraction (NCE) to catch the natural contour of non-rigid object. Different segment is labeled differently, the point belongs to the same segment gets the same label, so we can get the boundary points by the mutation of label.



**Fig. 4.** The object model: (a) The points of the contour are labeled, (b) the points located in the internal area of the contour are labeled with the same label

According to SLIC algorithm, only if there are more than one point among 8-neighbourhood (8 red points) of the yellow point whose label is different with it, it can be regard as boundary point. As shown in figure 3(a), the yellow point is initially labeled as a boundary point. Nevertheless, if we follow the rule of SLIC, this point will not be judged as a boundary point, therefore the natural contour extracted by NCE will not be a closed curve. For solving this problem, we improved the algorithm shown in figure 3(b), if there is one or more than one point among 4-neighbourhood (4 green points) of the yellow point whose label is different with it, it will be judged as the boundary point.

As is shown in figure 3(c), the non-rigid object is constituted by different segments and the natural contour of the non-rigid object is constituted by some boundary points of the segments. Labeling the segments of the non-rigid object with the same label, the closed natural contour of the object can be extracted by the principle proposed above. The result is shown in figure 3(d).

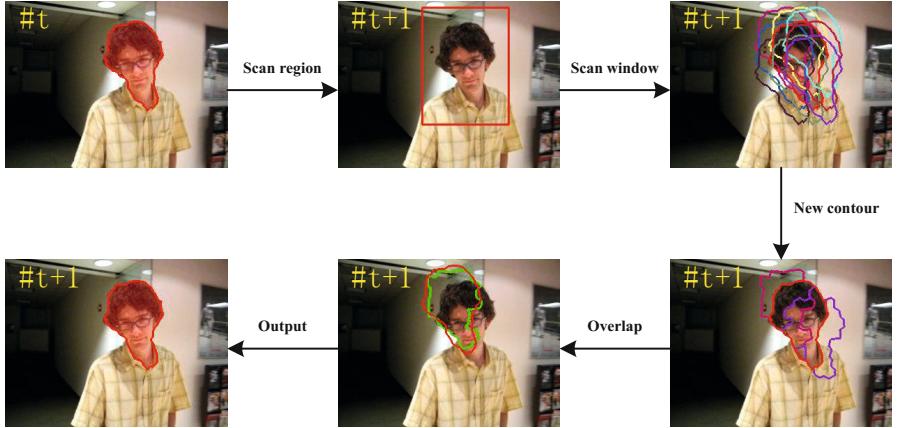
## 2.2 Detection Based on Flexible Scanning Window

We can catch the initial natural contour  $C$  of the object in frame  $t$ -th by the proposed method NCE. The internal area of  $C$  (including the points on the contour) is marked as  $A_C$ . Supposed that there are  $k$  points on the contour  $C$ , we can use  $C_i(i = 1, 2 \dots k)$  to represent each point on the contour, and the coordinates of each point are  $C_i(x, y)$ . Then the rectangular scanning region  $Bsc$  of frame  $(t+1)$ -th can be determined by the enclosing rectangle  $Ben$  of the area  $A_C$ ,

$$Bsc = Rect(x_{min}, y_{min}, x_{max} - x_{min}, y_{max} - y_{min}), \quad (1)$$

$$Ben = Rect(Bsc.x - L, Bsc.y - L, Bsc.width + 2L, Bsc.height + 2L), \quad (2)$$

where  $L$  is the increasing length of each edge when  $Bsc$  expanded into  $Ben$  and  $x_{min}, x_{max} \in \{x|C_i(x, y)\}$ ,  $y_{min}, y_{max} \in \{y|C_i(x, y)\}$ . If we set the scanning



**Fig. 5.** The block diagram of detection framework

step as  $stp$  on the directions of  $x$  and  $y$  axis respectively,  $num$  is the number of flexible scanning window  $SW$  which is computed as

$$num = \left( \frac{2L}{stp} + 1 \right)^2. \quad (3)$$

Then, we can get each point  $P_{ji}(j = 1, 2...num)(i = 1, 2...k)$  and its coordinates  $P_{ji}(x, y)$  of these scanning windows  $SW_j(j = 1, 2...num)$ . For any of  $j$ , the points  $P_{ji}(x, y)$  of contour  $SW_j$  (figure 4(a)) and the points of region  $A_{SW_j}$  are all labeled with  $L_{fore}$ , the rest points are labeled with  $L_{back}$ , we can get the object model  $M_j$  which is shown in figure 4(b).

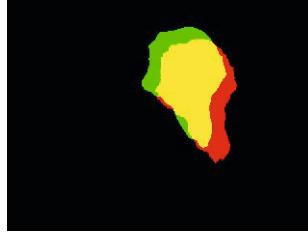
Combining with the segment information of frame  $(t+1)th$ , the set of seed points (each segment has a cluster center shown as green points on figure 3(c) which we called seed points) located in the internal area of the contour (the region in white) is computed as

$$U_j = \{(x, y) | M_j(x, y) = L_{fore}, (x, y) \in S\}, \quad (4)$$

where  $M_j(x, y)$  is the label of point  $(x, y)$  and  $S$  is the set of seed points in frame  $(t+1)th$ . A new contour  $NC_j$  can be calculated by NCE with  $U_j$  as input. In the end, the output contour  $Cout$  can be selected by the overlap between  $A_{SW_j}$  and  $A_{NC_j}$ , it is computed as

$$Cout = NC_j = \max_{j=1}^{num} \{2(A_{NC_j} \cap A_{SW_j}) - (A_{NC_j} \cup A_{SW_j})\}. \quad (5)$$

In order to understand our detection algorithm more intuitively, the block diagram of our detection framework is shown in figure 5. What we have done on every step is shown on the framework clearly. First, we use the natural contour



**Fig. 6.** The schematic of overlap computing: the region in green is ground truth and the region in red is the result of different algorithms (our detector or HT)

of frame  $t$ -th as input of our detector. Second, we figure out the rectangular region which we need to scan and construct the flexible scanning windows. Then, we compute the corresponding new contour of every scanning window and the overlap between them (two curves in 5th image of figure 5). Finally, according to the equation defined by overlap, we can get the output contour of frame  $(t+1)$ th.

### 3 Experiments

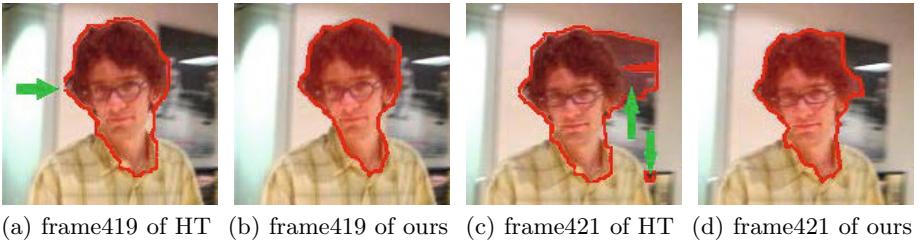
Except for HoughTrack, there is few detection or tracking algorithm whose result is shown with a natural contour, so we chose HT algorithm for comparison. We used an Intel Core i5-4250U (1.3GHz) machine with Ubuntu 10.04 to run our detector and HoughTrack. To prove the reliability of our detector, we use the test sequence (David) of a publicly available standard tracking dataset by Babenko et al.[13]. Besides, we use other two test sequences, one (Diving) is from HT[7] and the other (Walking) is ourselves. From the results shown in figure 9, we can see that the output contour computed by our detector fit the truly natural contour better than the results of HT obviously.

For quantitative analysis, we designed the flowing contrast experiment. As shown in figure 6, the region in green  $A_g$  is called ground truth (human segmented images in this case), the region in red  $A_r$  is computed by algorithm (HT or our detector) and the region in yellow (green+red) is overlap region. The overlap criterion in tracking is mostly defined as  $(A_g \cap A_r)/(A_g \cup A_r)$ , it may lead to a wrong judgement when  $A_g \subsetneq A_r$ , so we use  $2(A_g \cap A_r) - A_r$  to describe the overlap, and it is computed as

$$\text{overlap} = \frac{2(A_g \cap A_r) - A_r}{A_g} \times 100\%. \quad (6)$$

**Table 1.** The results of the contrast experiment

Algorithm/Aspect	Average overlap	Average running time
Our detector	<b>93.41%</b>	<b>146ms</b>
HoughTrack	<b>76.25%</b>	<b>633ms</b>



(a) frame419 of HT (b) frame419 of ours (c) frame421 of HT (d) frame421 of ours

**Fig. 7.** Comparison on details of the results: (a) is HT result of frame419 and (c) is HT result of frame421, the green arrows point out the defects of HT

The average overlap and processing time are shown on table 1. As what we can see, the average overlap of our detector is higher than the average overlap of HT, which shows that our detector has higher precision. And the average processing time of HT is more than 4 times longer than ours (including the time of gSLIC segmentation), which shows that our detector has stronger efficiency.

During Hough Tracker, it combines Hough Forests and Hough Voting to get the points located in object region, then, it uses a rough segmentation based on GrabCut to get the output contour. If it generates some wrong points when voting, it will lead a wrong segmentation which is shown in figure 7(c). Besides, the contour of HT is not a closed curve(figure 7(a)). But our detector will not generate those errors, we use gSLIC real-time segmentation to preprocess the frame sequence and get high quality segments (superpixels), then we design our detector based on superpixels instead of pixels. It not only reduces our calculation, but also provides us accurate and useful edge information. Combining with NCE algorithm, we can catch accurate natural contour (figure 7(b),(d)) of non-rigid object with very little time.

Just as mentioned in the paper[7], HT has defined a maximum object size that is used for background initialization of its segmentation algorithm, if the segmentation fails, it is not allowed to grow beyond this maximum scale. When we tested the Walking sequence, HT algorithm failed directly because of the size of the initial box shown in figure 8(a). The result of HT is seriously limited by the initial box, HT failed for most frames of Diving sequence (figure 8(c),(d)), even if the initial box is quite accurate (figure 8(b)).



**Fig. 8.** The limitations of the initial bounding-box in HT



**Fig. 9.** Results of three sequences. In each sequence, *Top row*: ours, *bottom row*: HT.

## 4 Conclusions

In this work, we present a superpixel-based detector that is able to avoid the problems brought by the bounding-box and handle the deformations of non-rigid object. By the combination of improved real-time superpixel segmentation algorithm, object-oriented natural contour extraction techniques and natural contour based detection method, we are able to track the accurate natural contour of non-rigid objects in videos. This will provide a reliable calculation basis for high-level computer vision tasks such as recognition and behavior analysis. We design our detector based on superpixels instead of pixels, which can improve the efficiency of our method. And we make full use of the superpixel boundary

information to extract natural contour of the object, then we construct flexible scanning windows with extracted contour. Therefore, our detector is able to perform such a challenging task. Experimental results show that the natural contour of our detector fit the truly natural contour better than the result of HT obviously. The average overlap of our algorithm achieves *93.41%* while HT is only *76.25%*. And our average running time is only *146ms* while HT needs *633ms*. In future work, we plan to combine our detector with tracking-by-detection mode and implement the whole approach on a Graphics Processing Unit (GPU). That would further improve the overall performance of our approach.

## References

1. Lascio, R.D., Foggia, P., Percannella, G., et al.: A real time algorithm for people tracking using contextual reasoning. *Computer Vision and Image Understanding* (S1077-3142) 117(8), 892–908 (2013)
2. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(S0162-8828) 34(7), 1409–1422 (2012)
3. Zhang, K., Zhang, L., Yang, M.-H., Zhang, D.: Fast Tracking via Spatio-Temporal Context Learning, vol. abs/1311.1939 (2013)
4. Wang, W., Nevatia, R.: Object Tracking Using Constellation Model with Superpixel. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 191–204. Springer, Heidelberg (2013)
5. Nejhum, S., Rushdi, M., Ho, J.: Visual Tracking Using Superpixel-Based Appearance Model. In: Chen, M., Leibe, B., Neumann, B. (eds.) ICVS 2013. LNCS, vol. 7963, pp. 213–222. Springer, Heidelberg (2013)
6. Yuan, Y., Fang, J., Wang, Q.: Robust Superpixel Tracking via Depth Fusion. *IEEE Transactions on Circuits and Systems for Video Technology* (accepted, 2013)
7. Martin, G., Peter, R., Horst, B.: Hough-based tracking of non-rigid objects. In: Proc. ICCV (2011)
8. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: ICCV (2011)
9. Alex, L., Adrian, S., Kiriakos, N., et al.: PixelTrack: a fast adaptive algorithm for tracking non-rigid objects. In: ICCV (2013)
10. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Ssstrunk, S.: SLIC Superpixels. Technical report (2010)
11. Levinstein, A., Stere, A., Kutulakos, K., Fleet, D., Dickinson, S., Siddiqi, K.: Turbopixels: Fast superpixels using geometric ows. *PAMI* (2009)
12. Ren, C.Y., Reid, I.: a real-time implementation of SLIC superpixel segmentation. Technical report. University of Oxford, Department of Engineering Science (2011)
13. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. In: Proc. CVPR (2009)

# An Improved Multipitch Tracking Algorithm with Empirical Mode Decomposition

Wei Jiang, Wenju Liu\*, Yingwei Tan, and Shan Liang

NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China  
`{wjjiang, lwj, ywtan, sliang}@nlpr.ia.ac.cn`

**Abstract.** Multipitch tracking is beneficial for speech separation, audio transcription and many other tasks. In this paper, we greatly improve a state-of-the-art multipitch tracking algorithm. While the amplitude and individual peak positions of autocorrelation function (ACF) were used in previous algorithms, a novel feature based on the average frequency of each time-frequency (T-F) unit is proposed in this paper. This feature is computed by an empirical mode decomposition (EMD) method. Then it is utilized to form the conditional probabilities in the hidden Markov model (HMM) given a pitch state of each frame, and finally the most likely state sequence is searched out. Quantitative evaluations show that the novel feature is more effective, and our algorithm significantly outperforms the previous one.

**Keywords:** multipitch determination algorithm, empirical mode decomposition, instantaneous frequency, HMM tracking.

## 1 Introduction

Pitch determination algorithms are very useful for many speech and audio signal processing techniques, such as audio retrieval and classification, tone recognition in Mandarin speech and co-channel speech separation. Well performed algorithms have been developed to determine a single pitch track for a clean speech utterance or its mixture with broad band noise. However, when the background noise has harmonic structure (e.g. music or speech of another person) or room reverberation exists, extracting both of the two pitch contours is very challenging.

Several cochleagram-based multipitch tracking algorithms have been proposed in previous studies. Alain de Cheveigné and Hideki Kawahara [1] brought with a temporal cancellation method for multiple period estimation. Wu *et al.* [2] modeled the probability of a cochleagram channel to support a candidate pitch by the relative time lag between the pitch period and its closest autocorrelation peak. Klapuri [3] incorporated an iterative spectral subtraction method with an auditory front end. More recently, Jin [4] utilized the amplitude of ACF to construct a new salience function, and much better results were obtained.

---

\* Corresponding author.

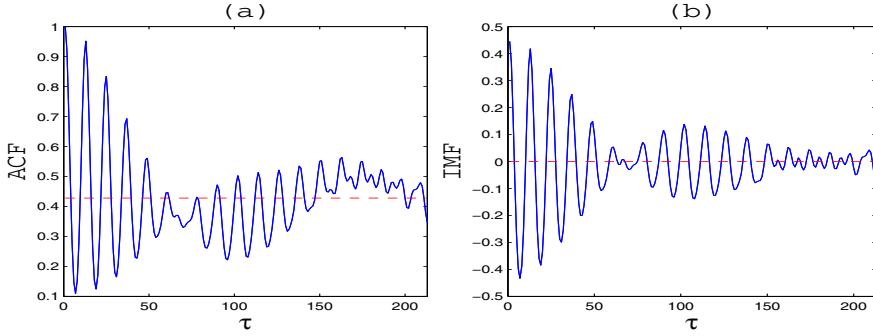
In these algorithms, a gammatone filterbank was used as an auditory front end, which has high frequency resolution in low channels and low resolution in high channels. Thus, high frequency channels usually contain multiple harmonics, and they are called unresolved channels. In channels of this kind, envelopes of autocorrelation functions (EACF) are utilized explicitly or implicitly in the above algorithms. This is reasonable because it has been proved that some auditory neurons are sensitive to the periodicity in the envelope of sounds [5]. In fact, EACFs of unresolved channels reflect the amplitude modulation (AM) phenomenon, thus they usually have peaks at fundamental periods in single pitch situation. However, we found that this is not always the case when two speakers are talking simultaneously, because different harmonics of different speakers may locate in the same channel, resulting in the mismatch of the global peak with neither of the two underlying pitches. Besides, energy disturbance such as room reverberation also leads to fluctuations of ACF/EACF amplitude and peak positions. Both conditions would cause the occurrence of errors in multipitch tracking. In this paper, we propose a novel feature which is based on the average frequency of each T-F unit. This feature is more stable and reliable than previous features (ACF/EACF amplitude and peak positions) under noisy and reverberant conditions, i.e. frequency is more stable than amplitude. This is the most important reason why great improvements is obtained.

The rest of this paper is organized as follows. Section 2 and 3 describe the feature extraction and multipitch tracking respectively. Results and comparisons are given in Section 4, followed by a conclusion section, where the relationship of our algorithm to prior work is also discussed.

## 2 Feature Extraction

First, the input signal  $x(t)$  is decomposed into 128 channels by a gammatone filterbank whose center frequencies range from 80 Hz to 3 kHz in a quasi-logarithmical way. The bandwidth of each filter is set by equivalent rectangular bandwidth (ERB). The output of each filter channel is further transformed into neural firing rate by Meddis model [6]. In each channel, the window length is set to be 20 ms with 10 ms shift, and a time-frequency representation of the original signal called *cochleagram* is obtained. We use  $u(c, m)$  to denote a T-F unit for frequency channel  $c$  and time frame  $m$ . The autocorrelation function for each T-F unit is then computed, and a representation called *correlogram* [7] is thus formed. In this section, we extract a correlogram-based feature which is computed by average frequency of each T-F unit for pitch tracking.

**Frequency Matching Function.** As described in Section 1, the new feature is based on average frequency which represents the mean of instantaneous frequencies within each T-F unit. However, filter outputs of unresolved channels are usually not narrowband signals. So how to define their instantaneous frequencies becomes a key problem. To define a meaningful instantaneous frequency, Huang *et al.* [8] proposed the concept of intrinsic mode function (IMF). An IMF is not



**Fig. 1.** Illustration of ACF and its first IMF. The red dashed line crosses the mean. (a) ACF; (b) IMF.

restricted to a narrowband signal, and it can be both amplitude and frequency modulated. Besides, IMFs can be obtained from complicated signals through a sifting process, which is also called an empirical mode decomposition (EMD) method. Thus, both resolved and unresolved channels have meaningful instantaneous frequencies. In fact, the average frequency of the first IMF in a unit represents that of the dominant speech, according to the principle of auditory masking. So we define the the average frequency of a unit based on its first IMF. Further, we could more robustly compute the average frequency of each unit by the zero crossing rate of the first IMF of ACF in each unit (not of the filter outputs themselves), as the phase of each unit is not concerned. This can be described by the following formula

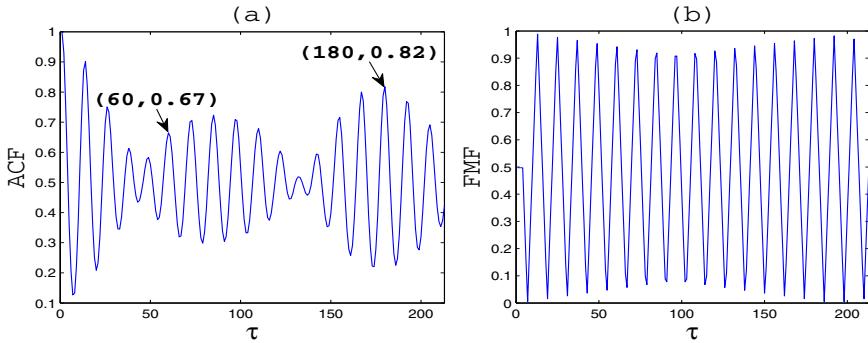
$$\bar{f}(c, m) = 0.5 \cdot \frac{N(c, m) - 1}{\tau_L(c, m) - \tau_F(c, m)} \cdot f_s \quad (1)$$

where  $\tau_F$  and  $\tau_L$  are the first and the last zero crossing point of the first IMF of ACF respectively.  $N(c, m)$  is the total number of zero crossing points in this unit, and  $f_s$  is the sampling frequency. Fig. 1 shows an example of an ACF and the first IMF of it. It can be seen from the figure that the IMF is symmetric with respect to the local zero mean, so its average frequency can be computed by formula (1).

Then, the supporting degree of the current channel for a candidate pitch period  $\tau$  can be modeled by

$$F(c, m, \tau) = 1 - 2 \cdot |\bar{f}(c, m) \cdot \tau - \text{int}(\bar{f}(c, m) \cdot \tau)| \quad (2)$$

Here,  $\tau$  corresponds to a candidate pitch period, and  $\text{int}(\cdot)$  returns the closest integer. It can be seen that  $F(c, m, \tau)$  ranges from 0 to 1. When  $\bar{f}(c, m)$  is exactly a multiple of the candidate pitch frequency,  $F(c, m, \tau)$  equals 1, which means that the current unit completely support the corresponding candidate pitch period. Generally speaking, the closer the  $\bar{f}(c, m)$  of a unit is to a multiple of the candidate pitch frequency, the closer the  $F(c, m, \tau)$  is to 1. In this sense, we may refer to this function as frequency matching function (FMF) of  $u(c, m)$ .



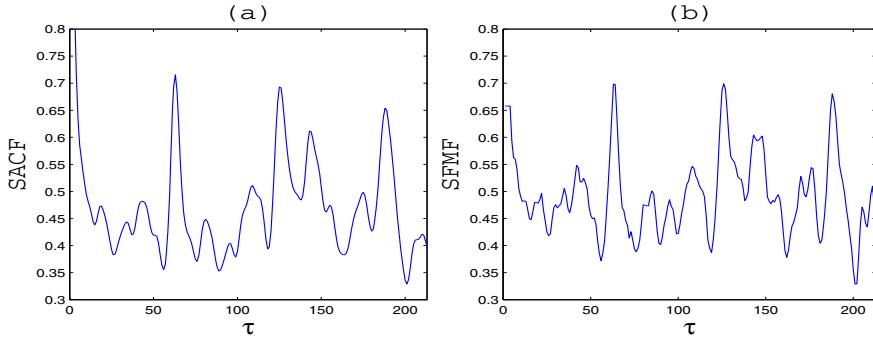
**Fig. 2.** Illustration of ACF and FMF. (a) ACF; (b) FMF.

## 2.1 Properties of FMF

In this subsection, we illustrate the advantages and disadvantages of FMF, compared with ACF. FMF focuses on the frequency aspect of a unit, and the effectiveness of it can be illustrated by Figure 2. Figure 2 shows us an example of ACF and its corresponding FMF. From Fig.2.(a), it can be seen that the auto-correlation function (ACF) has been severely affected by AM effect. However, the maximum peak around the delay point of 180 (i.e. 90Hz) corresponds to neither of the two underlying pitch periods, which are 60 (i.e. 271 Hz) and 140 (i.e. 115 Hz). In fact, the fifth harmonic ( $271 \times 5 = 1355$  Hz) of the higher pitch and the eleventh harmonic ( $115 \times 11 = 1265$  Hz) of the lower pitch locate together in this channel, and their difference is just 90 Hz which causes the global AM peak around 180. This means that the EACF peak (or ACF amplitude) is not reliable in this channel. On the other hand, it can be seen from Fig.2.(b) that FMF has boosted the peaks of ACF onto the same level considering sampling accuracy, and the adverse effect of AM has been removed in the sense that the real peak at 60 has been boosted. Similar adverse effect of amplitude fluctuations brought about by room reverberation could also be removed in this way.

There are two questions need to be clarified. First, some may wonder whether this boosting process also bring too many pseudo peaks. We found that this is not the case, because the summation process of the FMFs on all the channels of a frame will enable the pseudo peaks cancel each other. In fact, the sum auto-correlation function (SACF) and the sum frequency matching function (SFMF) are very similar in other normal situations. An example can be seen in Fig. 3. Second, there were channel selection processes based on cross correlation of neighboring channels in previous algorithms. Some may wonder whether this channel selection scheme excludes all corrupted channels. We found that this method not always work, because these channels sometimes dominate several adjacent channels, and their cross correlations are too high to be excluded.

As for the disadvantages of FMF, we can again look at Figure 1. Fig.1.(b) shows us that the instantaneous frequency of this unit rises from the delay point of 150 as the zero crossing rate becomes higher. In fact, this frame is a transitional



**Fig. 3.** Illustration of SACF and SFMF. (a) SACF; (b) SFMF.

frame that is not stable. So the average frequency of this unit will no longer correspond to the dominant pitch. On the other hand, Fig.1.(a) shows us that the amplitude of the ACF still reaches a peak around the delay point of 160 which corresponds to the dominant pitch period.

In summary, FMF is a new feature that focuses on the average frequency of a unit. It can avoid the adverse effect of amplitude modulation (AM) brought about by intrusion and reverberation, but it is not immune to frequency modulation (FM). However, adverse AM effect is more responsible for the occurrence of pitch tracking errors under multipitch and reverberant conditions. So better results can be expected by using FMF instead of ACF.

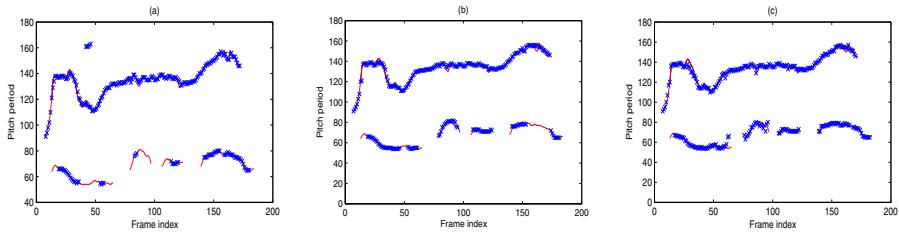
### 3 Multipitch Tracking

In this paper, we aim to track up to two pitches simultaneously. As in [4], the pitch state space can be defined as a union space of three subspaces known as zero-pitch space, one-pitch space and two-pitch space

$$\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2 \quad (3)$$

The pitch state of each frame is treated as a hidden state, and the candidate pitch range is [80, 500 Hz] in this paper.

In order to use the Viterbi algorithm to get the optimal pitch track, two kinds of probabilities have to be decided: transition probability and conditional probability (or observation probability). For transition probability, we use the same model and parameters as in [2], while the construction of conditional probability is the same as in [4] except that ACFs are replaced by our new features (FMFs). Finally, the Viterbi algorithm is adopted to search for the optimal pitch state sequence. To reduce the computational burden, we restrict the searching area to the pitch states that scatter around peak values of conditional probabilities and a pruning process is also included.



**Fig. 4.** Pitch tracking results for a mixture of one male and one female utterance. (a)-(c) plot detected pitch contours from Wu et al.’s , Jin et al.’s and our algorithm respectively. The red solid lines indicate the reference pitch tracks. The “\*” tracks represent the estimated pitch contours.

## 4 Experimental Results and Discussion

To verify the effectiveness of our algorithm, we compared our algorithm with two previous multipitch tracking algorithms proposed by Wu et al. [2] and Jin et al [4]. Besides, we evaluated it on two different data sets. To evaluate the performance of our algorithm quantitatively, we used the metrics proposed in [2]. Generally, there are two kinds of errors in multipitch tracking. First, total error  $E_t$  includes transition error and gross error. Transition error is denoted by  $E_{x \rightarrow y}(x, y = 0, 1, 2)$ , which happens when the pitch state of a frame is regarded as coming from the state space  $\Omega_y$  when it actually comes from  $\Omega_x$ . Gross error  $E_g$  is proportional to the number of frames where the detected pitch periods differs from the true pitch periods by more than 20 percent. Second, fine error  $E_f$  is defined as the average difference between the detected pitch periods and the true pitch periods for those frames that have no gross errors. All the reference pitch contours are extracted from clean utterances with Praat [12] and manual modifications.

### 4.1 Evaluation on Cooke’s Set

The first data set contains 30 speech mixtures. Each speech mixture contains about 160 frames, and the total frame number in the evaluation corpus is 5017. The speech utterances are selected from Cooke’s corpus [9]. These speech sentences include 10 target speech sentences of two male speakers and 3 intrusion sentences that are spoken by two female speakers and one male speaker.

Table 1 gives the pitch detection results of the two previous algorithms and our algorithm on the first evaluation corpus described above. It can be seen from the table that Jin et al.’s algorithm is able to do a better job in balancing between one- and two-pitch hypotheses than Wu et al.’s, and our algorithm does even better than Jin et al.’s with respect to almost all the criteria. On the whole, our total error rate is lowered by about 3 percent than Jin et al.’s algorithm. Fig. 4 shows an example of the pitch track contours detected by these three algorithms. As can be seen from the figure, transition errors account for the majority of the

**Table 1.** Error rates (in %) on Cooke's set

system	Wu <i>et al.</i>	Jin <i>et al.</i>	Proposed
$E_{0 \rightarrow 1}$	1.11	0.26	0.22
$E_{0 \rightarrow 2}$	0.02	0.08	0.00
$E_{1 \rightarrow 0}$	1.31	1.29	1.30
$E_{1 \rightarrow 2}$	0.64	1.59	1.34
$E_{2 \rightarrow 0}$	0.04	0.02	0.02
$E_{2 \rightarrow 1}$	25.29	18.96	16.50
$E_g$	0.20	0.72	0.86
$E_t$	28.61	22.93	<b>20.24</b>
$E_f$	0.65	0.69	0.67

**Table 2.** Error rates (in %) on Jin's set

$T_{60}(s)$	0.0		0.3		0.6	
	system	$E_t$	$E_f$	system	$E_t$	$E_f$
Wu <i>et al.</i>	34.39	0.98	45.24	1.30	57.81	2.03
Jin <i>et al.</i>	31.40	0.80	40.08	1.15	44.80	1.55
Proposed	<b>27.53</b>	0.81	<b>36.79</b>	1.22	<b>41.98</b>	1.85

total errors of pitch tracking. Specifically, the undetected error  $E_{2 \rightarrow 1}$  is evident in the multipitch tracking results, but this kind of error is obviously lowered by our algorithm.

In fact, the salience function proposed in [4] is based on the amplitude of ACF. However, as described in Section 1, this is not so reliable as frequency-based features under multipitch and reverberant conditions. So our algorithm obtained better results.

## 4.2 Evaluation on Jin's Set

Further, we evaluated our algorithm on another data set. This set is a much larger data set proposed in [4], which contains 50 anechoic speech mixtures as well as 300 mixture utterances with reverberation. Specifically, These speech sentences include 10 target speech sentences of five male speakers and five females, and 5 intrusion sentences that are spoken by three female speakers and two male speakers. For reverberant recordings, two reverberation time ( $T_{60}$ ) at 0.3 and 0.6 s are included, and each has three configurations.

Table 2 gives the multipitch detection results of the three algorithms in different reverberant conditions. For the sake of brevity, only total and fine errors ( $E_{tl}$  and  $E_{fn}$ ) are reported here. As can be seen from the table, our algorithm obtained a consistently better performance in both anechoic and reverberant conditions, which again verifies the effectiveness of our feature in lowering the total error. Compared with Jin's algorithm, the decreases of the total error rates are respectively 3.87, 3.29 and 2.82 percent in the three conditions. It seems that improvements are more difficult to achieve under highly reverberant conditions.

Meanwhile, there is some increase in the fine error rates compared with Jin's algorithm, but it is still lower than the results of Wu's algorithm. As illustrated in Section 2, FMF feature is not immune to FM, this is the reason for the minor rise of fine error in our results.

To summarize this section, our algorithm makes obvious improvements than Jin *et al.*'s algorithm. Similar results are obtained on two different kinds of corpora under anechoic as well as reverberant conditions. This section verifies the effectiveness of our feature proposed in Section 2.

## 5 Conclusions

This paper has proposed an improved HMM-based multipitch tracking algorithm for speech mixtures of two persons who are talking simultaneously. The utilization of frequency-based feature (FMF), which is calculated through an EMD method, avoids the adverse effect of AM in high frequency channels while SFMF retaining similar characteristics to SACF. Finally, quantitative evaluations show the consistent improvements of our algorithm compared to a state-of-the-art one under anechoic as well as reverberant conditions.

In previous studies, A. de Cheveigné *et al.* proposed a time-domain cancellation model [13] and then combined this model with an auditory front end [1]. Latter, Wu et al. [2] extended their work to a HMM-based joint search framework to determine the two pitch tracks simultaneously, and much better results were achieved. More recently, Jin *et al.*'s work obtained the best results of them. While the present study is related to recent HMM tracking approaches, it proposes a new feature, which captures the *frequency* aspect of T-F units. This feature is verified to be effective in resisting interferences and reverberation. On the contrary, the work by Wu *et al.* utilized the *peak positions* of EACF in high channels, and Jin *et al.* [4] used the *amplitude* of ACF to construct conditional probabilities. However, neither peak position nor amplitude of EACF/ACF is reliable in high channels under multipitch and reverberant situation. This is the reason why our algorithm performs better.

## References

1. Cheveigné, A., Kawahara, H.: Multiple period estimation and pitch perception model. *Speech Commun.* 27, 175–185 (1999)
2. Wu, M.Y., Wang, D.L., Brown, G.J.: A multipitch tracking algorithm for noisy speech. *IEEE Trans. Speech and Audio Processing* 11, 229–241 (2003)
3. Klapuri, A.: Multiple fundamental frequency estimation by summing harmonic amplitudes. In: Proc. Int. Conf. Music Inf. Retrieval (ISMIR), pp. 216–221 (2006)
4. Jin, Z.Z., Wang, D.L.: HMM-based multipitch tracking for noisy and reverberant speech. *IEEE Trans. Audio, Speech, Lang. Process.* 19, 1091–1102 (2011)
5. Schnupp, J., Nelken, I., King, A.: Auditory Neuroscience: Making Sense of Sound, pp. 128–129. MIT Press, Cambridge (2011)
6. Meddis, R.: Simulation of auditory-neural transduction: Further studies. *J. Acoust. Soc. Amer.* 83, 1056–1063 (1988)

7. Slaney, M., Lyon, R.F.: On the importance of time a temporal representation of sound. In: Visual Representations of Speech Signals, pp. 95–116. Wiley, New York (1993)
8. Huang, N.E., Shen, Z., Long, S.R., Wu, M.L., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. Roy. Soc. London A 545, 903–995 (1998)
9. Cooke, M.P.: Modeling Auditory Processing and Organization. Cambridge University, U.K (1993)
10. Zwicker, E.: Psychoacoustics. Springer, New York (1982)
11. Liu, W.J., Zhang, X.L., Jiang, W., et al.: Monaural voiced speech segregation based on elaborate harmonic grouping strategies. Sci. China. Inf. Sci., 2471–2480 (2011)
12. Boersma, P., Weenink, D.: Praat: Doing Phonetics by Computer (2004), <http://www.praat.org>
13. Cheveigné, A.: Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Am., 3271–3290 (1993)

# Robust Appearance Learning for Object Tracking in Challenging Scenes

Jianwei Ding<sup>1</sup>, Yunqi Tang<sup>1</sup>, Huawei Tian<sup>1</sup>, and Yongzhen Huang<sup>2</sup>

<sup>1</sup> People's Public Security University of China, Beijing, China

<sup>2</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

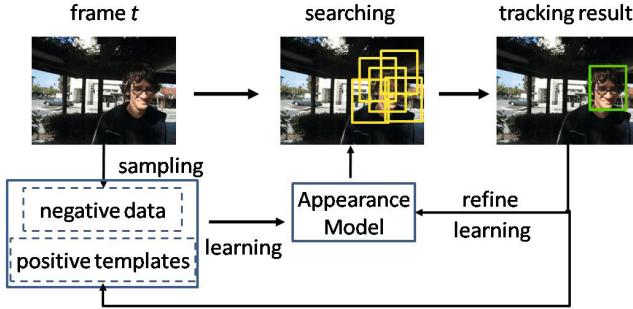
**Abstract.** This paper studies the appearance learning of object tracking in challenging scenes. We propose a new appearance modeling approach in the deep learning architecture for object tracking. Visual prior is learned from a large set of unlabeled images. Then it is transferred to the appearance model during tracking. Traditional trackers usually do tracking before updating at every input image. Drift may occur when there are complex appearance variations. We propose to update the appearance model before tracking. This can effectively prevent tracking failures when there are complex appearance changes. And the motion parameters estimation could be more accurate with the updated appearance model. Experimental results on challenging videos demonstrate the robustness and accuracy of the proposed algorithm compared with several state of the art approaches.

**Keywords:** object tracking, deep learning, challenging scenes, semi-supervised learning.

## 1 Introduction

Object tracking is a hot topic in computer vision. It is important for many visual applications, such as video surveillance, human-computer interaction, vehicle navigation and augmented reality. Although much progress has been made in recent years. And many tracking algorithms have been proposed [1]. It is still difficult to design a long term and robust tracking algorithm in challenging conditions.

Appearance modeling plays an important role in object tracking. It tries to learn an invariant representation of the object. Many modeling strategies have been proposed. And they can be divided into two categories: generative methods (e.g. [2], [3] and [4]) and discriminative methods (e.g. [5], [6] and [7]). A generative method treats tracking as searching for the region with the highest similarity to a pre-trained or online learned appearance model. In [2], an online subspace learning algorithm is proposed to incrementally update the appearance model during tracking. In [4], tracking by utilizing sparse representation is introduced to address the problem of large image corruptions. The generative methods often have better generalization performance when the size of training



**Fig. 1.** The flowchart of the proposed tracking algorithm

data is small, but have poor discriminative abilities when the background and object are similar. A discriminative method treats tracking as a binary classification problem between object and background. Thus it performs better to separate object from background. Grabner et al. [6] propose an online AdaBoost feature selection algorithm for tracking. Their method could run in real-time. In [5], online multiple instance learning is used instead of supervised learning to avoid drifting problem caused by incorrectly labeled training examples. As the discriminative methods are more robust to occlusion and strong variations, they are very popular in recent years.

At every input image, the location of the object is detected by the appearance model at first. Then the appearance model is online updated by the tracking result. This strategy of updating the appearance model after tracking is followed by most tracking algorithms. However, there is a gap between the appearance model and the object when there are fast or drastic appearance changes. Drift may easily occur and tracking would gradually fail.

In this paper, we propose a new discriminative tracker. Unlike traditional algorithms, our method updates the appearance model before tracking at every input frame. Fig. 1 shows the flowchart of the proposed method. First, negative samples with high confidence are collected from the input image. These samples are far away from the estimated location of the target at previous frame. Then the appearance model is updated with the negative samples and previous positive templates. The location of the object at current image is detected by the updated appearance model. Finally, the appearance model is refined by the tracking result. This strategy of learning before tracking can effectively prevent drift or tracking failure when there are sudden or fast appearance variations. Besides, location estimation can be more accurate because of using the updated appearance model in the tracking process.

We use the deep learning method to construct the discriminative appearance model. Without manual work of designing appearance features (e.g. Haar, color histogram and LBP), the deep learning method can automatically learn invariant representation of the target from raw data. Besides, visual prior can be learned from a large set of unlabeled images. Then it is transferred to the

appearance model during tracking. Similar strategy can be seen from [8]. The effectiveness of learning deep feature hierarchies have been demonstrated by a number of practical tasks, such as scene labeling [9], pedestrian detection [10], object classification [11].

The remainder of the paper is organized as follows. An overview of the proposed algorithm is introduced in the next section. The appearance model is described in Section 3. Experimental results and analysis are presented in Section 4. We draw conclusions in Section 5.

## 2 Overview of the Proposed Algorithm

Our tracking algorithm is based on the particle filter framework, which is able to deal with nonlinear/non-Gaussian motions. Thus object tracking is formulated as an inference task in a Markov model with hidden state variables [12]:

$$p(\mathbf{X}_t | \mathbf{O}_t) \propto p(\mathbf{o}_t | \mathbf{X}_t) \int p(\mathbf{X}_t | \mathbf{X}_{t-1}) p(\mathbf{X}_{t-1} | \mathbf{O}_{t-1}) d\mathbf{X}_{t-1} \quad (1)$$

where  $\mathbf{O}_t = \{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t\}$  is a set of observations and  $\mathbf{X}_t$  describes the motion parameters of the target at time  $t$ .  $\mathbf{o}_t$  is the observation vector at time  $t$  and  $\mathbf{o}_t \in \mathbb{R}^{d \times 1}$ . We use a bounding box with four parameters to describe the state variable:  $\mathbf{X}_t = (x_t, y_t, h_t, w_t)$ , where  $x_t, y_t, h_t, w_t$  denote  $x, y$  translation, height and width of the bounding box at time  $t$ . The appearance likelihood  $p(\mathbf{o}_t | \mathbf{X}_t)$  denotes the probability of the appearance data  $\mathbf{o}_t$  at state  $\mathbf{X}_t$ . And  $p(\mathbf{X}_t | \mathbf{X}_{t-1})$  represents the state transition probability between two consecutive frames.

Direct computation of the posterior probability  $p(\mathbf{X}_t | \mathbf{O}_t)$  is intractable, it is approximated by a finite set of  $N_s$  particle samples  $\{\mathbf{X}_t^k\}_{k=1}^{N_s}$  with importance weights  $\{w_t^k\}_{k=1}^{N_s}$ . Generally, the particle samples are drawn from an importance distribution  $q(\mathbf{X}_t | \mathbf{X}_{1:t-1}, \mathbf{O}_t)$ , and the weight of each sample is updated as

$$w_t^k \propto w_{t-1}^k \frac{p(\mathbf{o}_t | \mathbf{X}_t^k) p(\mathbf{X}_t^k | \mathbf{X}_{t-1}^k)}{q(\mathbf{X}_t | \mathbf{X}_{1:t-1}, \mathbf{O}_t)} \quad (2)$$

To avoid degeneration, the particles are resampled to generate a set of new equally weighted samples according to their importance distribution. In the case of the bootstrap filter, the importance distribution  $q(\mathbf{X}_t | \mathbf{X}_{1:t-1}, \mathbf{O}_t) = p(\mathbf{X}_t | \mathbf{X}_{t-1})$ , and the weights are obtained by normalizing the appearance likelihood  $p(\mathbf{o}_t | \mathbf{X}_t)$ .

The tracking process is governed by the motion model and the appearance model. The motion model describes the state transition probability  $p(\mathbf{X}_t | \mathbf{X}_{t-1})$  between two consecutive states. The state transition probability can be calculated with a Gaussian distribution:

$$p(\mathbf{X}_t | \mathbf{X}_{t-1}) = N(\mathbf{X}_{t-1}, \Sigma) \quad (3)$$

where  $N$  denotes the Gaussian distribution,  $\Sigma = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_h^2, \sigma_w^2)$  is the diagonal covariance matrix.

## 2.1 The Appearance Model

The appearance model is composed of a discriminative classifier, which is able to compute the appearance likelihood  $p(\mathbf{o}_t | \mathbf{X}_t)$ . It updates itself by learning from current input image at first. Then the tracking process is governed by the updated appearance model. The details will be described in section 3.

At each time  $t$ , suppose there are  $N_s$  particle samples  $\{\mathbf{X}_t^k\}_{k=1}^{N_s}$ , and the corresponding cropped image patches are  $\{\mathbf{o}_t^k\}_{k=1}^{N_s}$ . Each image patch  $\mathbf{o}_t^k$  is classified by the classifier, and the confidence of which is  $p(y_t^k = 1 | \mathbf{o}_t^k)$ . The appearance likelihood  $p(\mathbf{o}_t | \mathbf{X}_t)$  is computed directly by normalizing the confidence:

$$p(\mathbf{o}_t^k | \mathbf{X}_t^k) \propto p(y_t^k = 1 | \mathbf{o}_t^k), k = 1 \dots N_s \quad (4)$$

And the weight of each particle is calculate by  $w_t^k = p(\mathbf{o}_t^k | \mathbf{X}_t^k)$ . Thus the optimal state at time  $t$  is estimated by finding the image patch with the biggest weight:

$$\mathbf{X}_t^* = \arg \max_{\mathbf{X}_t^k} p(\mathbf{o}_t^k | \mathbf{X}_t^k) \quad (5)$$

If the biggest weight of the particles is below a predefined threshold  $\tau$ , the appearance model, or the classifier is online updated again by the tracking result. The details of how to maintain the appearance model will also be described in the next section.

The flowchart is summarized in Algorithm 1. To reduce computation time, the first step in Algorithm 1 can be simplified. In the experiment, we obtain maximum confidence score of particles predicted by the appearance model. If it is below a predefined threshold  $\varsigma$ , the appearance model is updated before tracking. Usually we set  $\varsigma < \tau$ .

---

**Algorithm 1.** The flowchart of the tracking algorithm

---

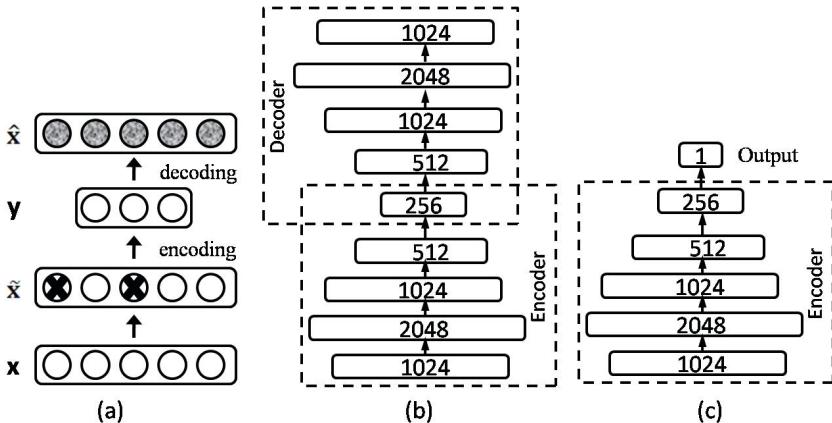
**Require:** The video frame at time  $t$

**Ensure:**

- 1: Extract negative samples from current image. Updates the appearance model with the negative samples and previous positive templates.
  - 2: Generate particle samples  $\{\mathbf{X}_t^k\}_{k=1}^{N_s}$  according to the motion model.
  - 3: For each particle, extract the corresponding image patch from the input image, and calculate the weight under the appearance model.
  - 4: The optimal state  $\mathbf{X}_t^*$  is estimated by Eq. 5.
  - 5: Store the image patch corresponding to the estimated state. If the weight of the image patch is below a predefined threshold  $\tau$ , incrementally update the appearance model.
- 

## 3 The Appearance Model

The appearance modeling plays an important role in object tracking. The representation of an object should be invariant to cluttered background, noise,



**Fig. 2.** (a) Basic denoising autoencoder. (b) Stacked denoising autoencoder. (c) The whole network of the appearance model.

illumination change, etc. We use the deep learning method to learn an invariant appearance model of the object. The deep learning method often involves unsupervised training with a large set of unlabeled data. Then the unsupervised trained model is fine tuned by some labeled samples. The training details are described below.

### 3.1 Unsupervised Pre-training

In the unsupervised pre-training stage, a generic appearance model is learned by training a stacked denoising autoencoder (SdA) with a large number of unlabeled images.

We use the CIFAR-100 dataset [13] for pre-training, which is a subset of 80 million tiny images dataset [14]. The CIFAR-100 dataset consists of 60000 32x32 color images in 100 classes, with 600 images per class. There are 50000 training images and 10000 test images. We use all of them as the training data. Each color image is converted to grayscale before training. Thus the dimension of each input sample is 1024.

A SdA consists of multiple layers of denoising autoencoder (dA) in which the outputs of each layer is wired to the inputs of the successive layer. The denoising autoencoder is a variant of the traditional autoencoder, which adds noise or corruption to the input. Thus it can prevent the appearance model from learning the identity function.

The basic denoising autoencoder is shown in Fig. 2 (a). The initial input vector is  $\mathbf{x}$ . And the partially destroyed version is  $\tilde{\mathbf{x}}$ . Then  $\tilde{\mathbf{x}}$  is mapped to a hidden representation  $\mathbf{y}$  through a deterministic mapping function:

$$\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad (6)$$

where  $\mathbf{W}$  is a weight matrix,  $\mathbf{b}$  is a bias vector, and  $s(\cdot)$  is a nonlinear function such as the logistic sigmoid function or the tanh function. The hidden representation  $\mathbf{y}$  is reconstructed to a new vector  $\hat{\mathbf{x}}$  by a mapping function:

$$\hat{\mathbf{x}} = s(\mathbf{W}'\mathbf{y} + \mathbf{b}') \quad (7)$$

where  $\mathbf{W}' = \mathbf{W}^T$  and  $\mathbf{b}'$  is a bias vector.

Suppose there are  $n$  training samples. The  $i$ th input sample is  $\mathbf{x}_i$ , and the reconstructed vector is  $\hat{\mathbf{x}}$ . Thus the loss function can be defined by:

$$L = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \lambda(\|\mathbf{W}\|_F^2 + \|\mathbf{W}'\|_F^2) \quad (8)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm. The first term is an average sum-of-squares error term, and the second term is a regularization term that forces to decrease the magnitude of the weights and helps to prevent overfitting.  $\lambda$  is the weight decay parameter, which controls the relative importance of the two terms.

To achieve sparse activities of the hidden units, we can specify a sparsity target  $\rho \ll 1$ . Let  $\hat{\rho}_j$  denote the empirical activation rate of the  $j$ th hidden unit. It can be computed by:

$$\hat{\rho} = \sum_{i=1}^n \mathbf{y}_i \quad (9)$$

An extra sparsity penalty term is added to the loss function by using the Kullback-Leibler (KL) divergence between the expected activation rate and the empirical activation rate:

$$KL(\rho \parallel \hat{\rho}) = \sum_{j=1}^{d'} \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j} \quad (10)$$

The parameters of the denoising autoencoder is obtained by minimizing the overall loss function:

$$(\mathbf{W}^*, \mathbf{b}^*, \mathbf{b}'^*) = \arg \min(L + \beta KL(\rho \parallel \hat{\rho})) \quad (11)$$

where  $\beta$  controls the weight of the sparsity penalty term.

The whole structure of the deep SdA is shown in Fig. 2 (b). The encoder is a 1024-2048-1024-512-256 neural network. The decoder is a network to recover the data from the code. In the first layer, we use overcomplete representation. To train parameters of the SdA, we use the greedy layer-wise approach. Each layer of the network is trained in turn and individually. Once the first  $k$  layers are trained, we can train the  $k+1$ -th layer as the representation of the  $k$ th layer can be used as the input of the  $k+1$ -th.

### 3.2 Online Appearance Modeling

After pre-training, the parameters of all layers are fine-tuned using backpropagation. A logistic regression layer is added to the encoder of the pre-trained SdA. The architecture of the whole appearance network is shown in Fig. 2 (c).

The initial appearance model of the object is generated by transferring the generic appearance model with the labeled bounding box of the object at the first frame. But it is not stable to train the initial appearance model with only one positive instance. Thus we generate more positive samples by synthesizing from the initial bounding box. First we select 10 bounding boxes that are very close to the initial bounding box. Then 100 warped versions of the selected bounding boxes are generated by geometric transformations (shift  $\pm 1\%$ , scale change  $\pm 1\%$ , in-plane rotation  $\pm 10^\circ$ ). Thus we obtain 100 positive samples. The negative samples are collected from the surrounding of the initial bounding box and the background. No geometric transformation is applied to the negative samples.

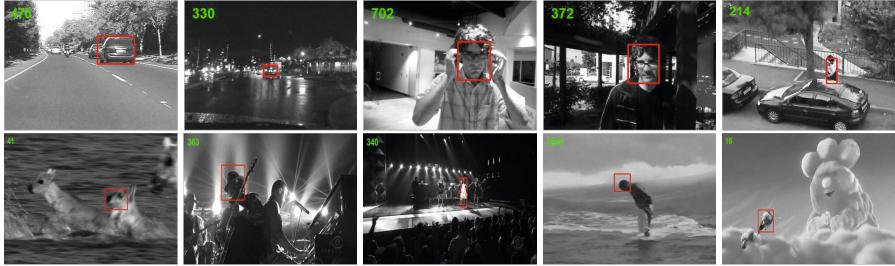
After initializing the appearance model, the tracker is ready for estimating locations of the object in subsequent frames. Unlike traditional tracking by detection methods, which do tracking before updating the appearance model, our method does tracking after updating the appearance model at every input frame. Negative training data is sampled from current input image. The positions of these samples are far away from the position of target at previous frame. Generally, the difference of the positions between two consecutive frames is small. So it is reliable to sample negative samples in this way. We use previously saved tracking results as the positive training data. Then the appearance model is updated with these training samples before tracking.

If the maximum confidence of the particles is below a predefined threshold  $\tau$  after tracking, the appearance model is updated with new training samples. The positive samples are previously stored tracking results and current tracking result. The negative samples are randomly selected around current estimated bounding box.

## 4 Experimental Results

To evaluate the performance of the proposed algorithm, we do experiments on challenging videos which are commonly used in the literatures and publicly available. And our method is compared with results reported in [15] which reported performances of several state-of-the-art trackers: DLT [15], MTT [17], VTD [18], MIL [5], a latest variant of L1T [19], TLD [16], and IVT [2]. Each tracker is initialized in the first frame and tracks the target up to the end. And the initial states are set to be the same in the first frame.

Some important parameters are described here. For unsupervised pre-training, the zero masked fraction of training samples is 0.001, the weight penalty parameter  $\lambda = 0.0001$ , the sparsity penalty parameter  $\beta = 0.0001$ , the sparsity target  $\rho = 0.05$  and the mini-batch size is 100. For online appearance updating, the



**Fig. 3.** Tracking results on ten challenging sequences

weight penalty parameter  $\lambda = 0.002$ , the sparsity penalty parameter  $\beta = 0.001$ , the sparsity target  $\rho = 0.05$  and the mini-batch size is 20.

The snapshots of the tracking results can be seen from Fig. 3. In the “car4” and “car11” sequences, the targets are cars moving on road. The challenging is the illumination change. In the “davidin” and “woman” sequences, the targets are faces which undergo drastic illumination and pose changes. In the “woman” sequence, the target is a walking woman which is severely occluded by the parked cars. In the “animal” sequence, the target is a fast moving animal which is blurred by abrupt motion. In the “shaking” and “singer1” sequences, the targets undergo complex and drastic illumination changes, which are very difficult to be tracked. In the “surfer” sequence, the pose of a surfer’s head changes drastically. The “bird2” sequence is also very challenging since the pose changes drastically and the target is occluded severely. Our method can track the target well in most of the sequences.

Two performance measures are used in the experiments: (1) center position error - the Euclidean distance between the tracked window centroid and the ground truth window centroid, and (2) success rate - if the tracked bounding box and the ground truth bounding box are  $\mathbf{A}$  and  $\mathbf{B}$  respectively, the overlap score is defined as  $S = \frac{\mathbf{A} \cap \mathbf{B}}{\mathbf{A} \cup \mathbf{B}}$ , where  $\cap$  and  $\cup$  represent the intersection and union of two regions respectively. The target is defined to be successfully tracked if the overlap score is bigger than 0.5.

The comparison results are summarized in Table 1. According to the performance metric of success rate, our method achieves the best in 4 sequences and the second best in 4 sequences. And according to the performance metric of center position error, our method achieves the best in 5 sequences and the second best in 3 sequences.

**Analysis.** It is difficult to track an object successfully in challenging scenes. The key is the appearance modeling. We use the deep learning method to learn an invariant appearance model of the object. This approach can automatically learn from raw data without extracting any invariant features. Besides, visual priors are learned from a large set of unlabeled images, which can also improve the tracking performances. Traditional trackers usually update the appearance model after tracking. This ignores the gap between the appearance model and

appearance data when there are fast appearance variations. Our approach updates the appearance model firstly, then tracks the object by the updated model. This can effectively prevent tracking failures when there are sudden or fast appearance changes. Besides, the motion parameters estimation can be more accurate with the updated appearance model.

**Table 1.** Comparison of 8 trackers on 10 video sequences. The first number denotes the success rate, while the number in parentheses denotes the center position error. The best and the second best performing methods are shown in red and blue respectively.

	Ours	DLT	MTT	VTD	MIL	L1T	TLD	IVT
car4	100(5.7)	100(6.0)	100(3.4)	35.2(41.5)	24.7(81.8)	30.8(16.8)	0.2(-)	100(4.2)
car11	100(1.1)	100(1.2)	100(1.3)	65.6(23.9)	68.4(19.3)	100(1.3)	29.8(-)	100(3.2)
davidin	78.1(6.5)	66.1(7.1)	68.6(7.8)	49.4(27.1)	17.7(13.1)	27.3(17.5)	44.4(-)	92.0(3.9)
trellis	76.1(7.0)	93.6(3.3)	66.3(33.7)	30.1(81.3)	25.9(71.7)	62.1(37.6)	48.9(-)	44.3(44.7)
woman	81.3(6.9)	67.1(9.4)	19.8(257.8)	17.1(133.6)	12.2(123.7)	21.1(138.2)	5.8(-)	21.5(111.2)
animal	100(4.9)	87.3(10.2)	88.7(11.1)	91.5(10.8)	63.4(16.1)	85.9(10.4)	63.4(-)	81.7(10.8)
shaking	96.2(10.1)	88.4(11.5)	12.3(28.1)	99.2(5.2)	26.0(28.6)	0.5(90.8)	15.6(-)	1.1(138.4)
singer1	100(3.1)	100(3.3)	35.6(34.0)	99.4(3.4)	10.3(26.0)	100(3.7)	53.6(-)	96.3(7.9)
surfer	85.3(5.6)	86.5(4.6)	83.8(6.9)	90.5(5.5)	44.6(14.7)	75.7(9.5)	40.5(-)	90.5(5.9)
bird2	60.2(18.4)	65.9(16.8)	9.2(92.8)	13.3(151.1)	69.4(16.3)	45.9(57.5)	31.6(-)	10.2(104.1)

## 5 Conclusions

In this paper, we address the challenge of appearance modeling in object tracking. A new appearance learning approach in the deep learning architecture is proposed for object tracking. Visual prior is learned from a large set of unlabeled images before tracking. Then it is transferred to the appearance model of the object. Traditional trackers usually do tracking before updating at every input image. Drift may occur when there are fast appearance variations. We propose to update the appearance model before tracking. This can effectively prevent tracking failures when there are complex appearance changes. And the motion parameters estimation is more accurate with the updated appearance model. To test the robustness and accuracy of the proposed tracker, several challenging public videos are used in the experiment. And the experimental results are compared with several state-of-the-art approaches. The results show that our method can work well in the challenging scenes.

**Acknowledgement.** This work is supported by the National High Technology Research and Development Program of China (863 Program) (No. 2013AA014604), the Fundamental Research Funds for the Central Universities (No. 2014JKF01116 and “Research on Efficiency Evaluation of Video Surveillance System Front and Object Detection”).

## References

1. Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A Survey. *ACM Computing Surveys* 38(4) (2006)
2. Ross, D.A., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, 125–141 (2008)
3. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence* 25(5), 564–577 (2003)
4. Mei, X., Ling, H.: Robust Visual Tracking using L1 Minimization. In: *IEEE Conf. on Computer Vision and Pattern Recognition* (2009)
5. Babenko, B., Yang, M.-H., Belongie, S.: Robust Object Tracking with Online Multiple Instance Learning. *IEEE Trans. Pattern Analysis and Machine Intelligence* 33(8), 1619–1632 (2011)
6. Grabner, H., Grabner, M., Bischof, H.: Real-Time Tracking via On-line Boosting. In: *Proc. British Machine Vision Conf.* (2006)
7. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: *Proc. Int'l Conf. on Computer Vision*, pp. 1515–1522 (2009)
8. Wang, Q., Chen, F., Yang, J., Xu, W., Yang, M.-H.: Transferring Visual Prior for Online Object Tracking. *IEEE Transactions on Image Processing* 21(7), 3296–3305 (2012)
9. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8), 1915–1929 (2013)
10. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian Detection with Unsupervised Multi-Stage Feature Learning. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2013)
11. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*, pp. 1106–1114 (2012)
12. Doucet, A., de Freitas, N., Gordon, N.: *Sequential Monte Carlo Methods in Practice*, New York (2001)
13. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
14. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(11), 1958–1970 (2008)
15. Wang, N., Yeung, D.-Y.: Learning a Deep Compact Image Representation for Visual Tracking. In: *Advances in Neural Information Processing Systems* (2013)
16. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-Learning-Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(7), 1409–1422 (2012)
17. Ahuja, N., Liu, S., Ghanem, B., Zhang, T.: Robust visual tracking via multi-task sparse learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2042–2049 (2012)
18. Kwon, J., Lee, K.M.: Visual Tracking Decomposition. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
19. Bao, C., Wu, Y., Ling, H., Ji, H.: Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1830–1837 (2012)

# Vehicle Recognition for Surveillance Video Using Sparse Coding

Shirong Zeng, Xin Niu, and Yong Dou

National Laboratory for Parallel and Distributed Processing,

National University of Defense Technology, Changsha, China

estbanroger@163.com, {niuxin,yongdou}@nudt.edu.cn

**Abstract.** This paper presents a vehicle recognition approach for a real transportation surveillance system using sparse coding. Comparison between sparse coding and conventional histogram of orientation gradient (HOG) has been studied. The results showed that the sparse coding learned feature is better than HOG feature in such vehicle recognition application. Experiments indicated that overlapping spatial pooling over the learned sparse codes can improve accuracy in a great deal.

**Keywords:** object recognition, sparse coding, vehicle surveillance.

## 1 Introduction

Vehicle Recognition is an essential application in intelligent transportation surveillance system. For recognition from video clips, extraction and representation of effective object features are of great importance.

In conventional image processing, most of the image features are extracted through a fixed flow path and represented by some predefined descriptors. Efficiency of some state-of-the-art image descriptors such as Histograms of Orientation Gradient (HOG) [1] and Scale-invariant feature transform (SIFT) [3] have been reported in many previous studies. One of the critical factors regarding the success of those descriptors is that they represent features with reduced sensitivity in the spatial or temporal variance. For example, HOG extracts moderately high level statistical features which are tolerant to translation. While SIFT resembles a hierarchical processing, which contributes to its insensitivity to scale and rotation.

Recently, the superiority of the data mined features has been noticed. One of the popular methods is sparse coding. Sparse coding tries to map the input data into a sparse feature space, and it has been proved that it is an efficient way to learn underlying structures of data. It is also found by Kai Yu [2] that, by a hierarchically organized sparse coder, the local dependencies can be learned. And this layer-wise structure is capable to extract mid-level features which are invariant to the individual change within the class. Although high performance on specific datasets has been reported, it seems not an easy way to train that layer-wise model with a large parameter space. On the other hand, by summarizing the local dependencies through certain spatial operations such as pooling

could also improve the recognition efficiency. However, to what degree such simplification could achieve, especially in comparison with some well performed image descriptors like HOG has not been well studied.

Therefore, in this paper, we present a classification approach using sparse coding and spatial pooling. This approach can be easily implemented to recognize vehicles from real surveillance system with acceptable accuracy. The proposed method was compared with HOG. Efficiency of the pooling operation in the approach was evaluated.

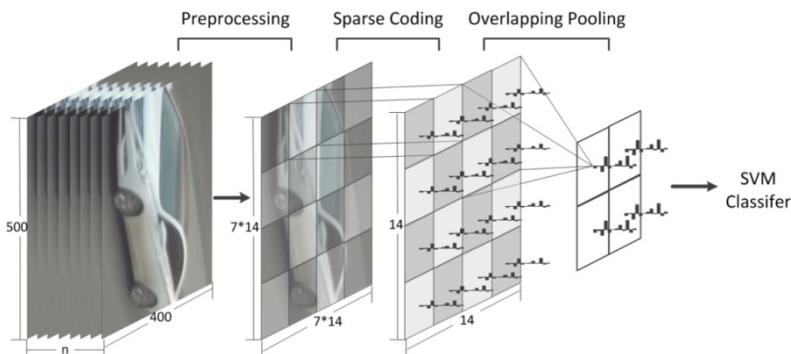
## 2 Methodology

### 2.1 Overview of the Proposed Method

In this section, we describe the frame of the proposed method. The dataflow is plotted in Figure 1.

This approach started from a preprocessing on the input images. Data preprocessing plays an important role in many deep learning algorithms. In many cases, results can be further improved after normalization and whitening data.

Once the preprocessing was finished, a sparse coding dictionary was trained. In order to construct the training set, a number of patches were randomly sampled from the images. Honglak Lee et al. [4] presented a practical method for coding and training. In terms of coding, images were partitioned into nonoverlapping patches, as the size as training patches. Each patch was then coded separately base on the dictionary. The coding procedure tried to obtain a group of sparse coefficients, meanwhile, reconstructed the original data as far as possible. The codes were pooled afterwards, and features from adjacent patches were pooled together and form a block feature vector. Finally the images were represented by the concatenated block-wise features. And a SVM was employed on such features to classify those images with typical vehicles.



**Fig. 1.** Framework of the method

## 2.2 Preprocessing

Surveillance video clips vary greatly in illumination. When using images from different illumination conditions without any preprocessing, acceptable results are usually hard to be achieved. Moreover, it's hard to avoid noises in surveillance systems. Generally, noises are in a higher range than the interested object in the frequency domain, thus can be reduced by a low-pass filter.

We carried out the preprocessing by a) rescaling the data along each dimension to  $[-1,1]$ . b) zero-meaning pixels in each image to unify illuminations. c) normalizing variance along each dimension by dividing its standard deviation. d) whitening the data using ZCA which also plays as a low-pass filter.

## 2.3 Sparse Coding

According to B.A.Olshausen's theory [5], an image can be represented in terms of a linear superposition of bases functions, that is

$$I(x, y) = \sum_{i=1}^p a_i \phi_i(x, y) \quad (1)$$

where  $\phi_i(x, y)$  is a base functions or dictionary elements, and  $a_i$  is the corresponding coefficient. Any images can be commendably reconstructed by a overcomplete dictionary whose rank is larger than  $p$ . To encourage a group of sparse coefficients, a  $l - Norm$  is imposed on the object function as

$$J_\Phi(x) = \sum_{(x,y)} \left\| I(x, y) - \sum_{i=1}^p a_i \phi_i(x, y) \right\|^2 + \lambda \|a_i\|_l \quad (2)$$

where  $\lambda$  is a constant to balance the reconstruction error and the sparsity of  $a$ .

## 2.4 Spatial Pooling

Pooling is a useful approach to achieve invariance to spatial change in most hierarchical or deep networks [6]. In a hierarchical model, it makes higher features more robust to noise and clutter. [6] Conventional pooling steps are implemented by a max or sum operation on the neighboring feature outputs by earlier stages in a nonoverlapping style. However, the nonoverlapping pooling can not efficiently explore the spatial dependencies between pooled windows. To improve the image object recognition with spatial relationships, the overlapping pooling scheme was applied in our method.

## 3 Experiment

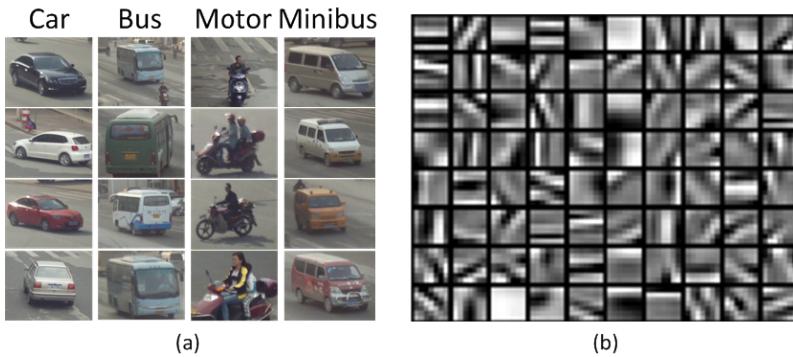
We evaluated the recognition accuracy of both HOG and sparse coding on the same dataset. A comparative analysis between pooled sparse codes and codes without pooling were made as well.

### 3.1 Experiment Setup

The dataset consists of 2520 images which equally distribute in 4 classes, i.e. Car, Bus, Motor and Minibus.  $7/8$  of images in the overall dataset are sampled to form a training set and the remaining as test set. Some samples from the dataset are illustrated in Figure 1(a).

All images are extracted from surveillance videos, and the cameras locate in several busy crossings. And the backgrounds are diverse and complicated. Compared with other datasets such as COIL-101 and MINIST, our vehicle dataset is specified and closer to real application. And classification between similar concepts, e.g. Bus and Minibus in our experiment, is of much difficulty.

The size of the extracted images is 500\*400, which is sufficient to cover the main bodies of most of the vehicles with few exceptions. To reduce the training time, images are resized to 108\*88 for HOG and 98\*98 for sparse coding.



**Fig. 2.** (a) Image samples from dataset. (b) Learned base functions of sparse coding.

The images are resized to 108\*88 for HOG. The HOG feature is computed in the similar way described in [1], except a modification for clipping edges of the gradient image. In a standard HOG procedure, it brings about invalid gradients in edges of the image due to the fact the filter  $[-1, 0, 1]$  would overlap the non-existent pixels outside the image. We cut them before voting, thus the cut image is actually 100\*80. It is then partitioned into non-overlapping cells of 7\*7. When normalizing contrast, the block is set to 2 cells x 2 cells, with a step of 1 cell x 1 cell.

For sparse coding, images are partitioned into non-overlapping patches of 7\*7.  $\lambda$  is set to an appropriate value so that the train process converges meanwhile the sparse term is larger than the error one. We make trials of different pooling configurations from 2x2 pooling window with 1x1 steps to 4x4 pooling window with 4x4 steps, as shown in Table 1.

### 3.2 Result and Discussion

Figure 1(b) shows the learned bases from vehicle picture patches. As the training process converges, we obtain basis functions that resemble Gabor filters with specified orientations.

When applied to recognition, our method obtains a modest accuracy on vehicle images (Table 1). Our experiment shows that the learned feature of sparse coding is more suitable than the manually designed HOG features for classification.

**Table 1.** Accuracy of recognition for raw images pixels, HOG, sparse coding without pooling and sparse coding with pooling

SC with pooling	window size	2x2		3x3			4x4			
	window step	1x1 (+)*	2x2 (-)	1x1 (+)	2x2 (+)	3x3 (-)	1x1 (+)	2x2 (+)	3x3 (+)	4x4 (-)
	accuracy	71.6%	71.3%	76.9%	74.0%	72.7%	76.6%	74.9%	71.1%	71.3%
SC without pooling	accuracy				56.2%					
HOG	accuracy				61.1%					
raw images	accuracy				23.8%					

\* (-) indicates nonoverlapping, and (+) is overlapped.

#### Overlapping

In terms of pooling, the experiment showed that pooling significantly affected the recognition result. We made trials of spatial pooling with various configurations (Table 1). Since original nonoverlapping patches might lead to spatial discontinuity in sparse coding features, we adopted the overlapping pooling scheme. It can be observed that the accuracy increased with the decreasing pooling window steps which demonstrates the benefits of the overlapping pooling in comparison with the nonoverlapping

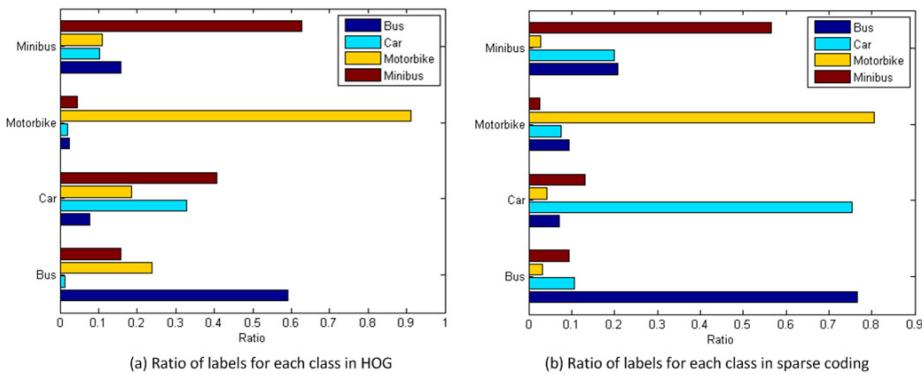
#### Window size

It can be also observed from Table 1 that, pooling window size also affected the result. On one hand, the mechanism of pooling implies neglect of objects details. On the other hand, the effects due to the noise could be reduced and the invariance could be well learned by the larger window. Appropriate window size balanced feature significance and its complexity. In our experiment, highest accuracy of recognition was achieved using  $3 \times 3$  sum pooling with  $1 \times 1$  steps.

#### Comparison on different classes

Classification results of HOG and sparse coding with configuration of  $4 \times 4$  window and  $(4,4)$  sum pooling is shown in Figure 3. It infers that in our dataset, motorbikes are always easier to be distinguished by both HOG and sparse coding. A lot of cars

are misclassified to minibus using HOG features, while a large number of minibuses are misclassified to cars or buses in sparse coding.



**Fig. 3.** Ratios of labels for each class in recognition

## 4 Conclusion

A vehicle recognition approach using sparse coding for real transportation surveillance videos has been proposed. The comparison between sparse coding and HOG illustrates that the learned feature by our approach is more feasible for vehicle identification. The recognition accuracy can be further improved through overlapping pooling.

## References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
2. Yu, K., Lin, Y., Lafferty, J.: Learning image representations from the pixel level via hierarchical sparse coding. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1713–1720. IEEE (2011)
3. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
4. Lee, H., Battle, A., Raina, R., et al.: Efficient sparse coding algorithms. Advances in Neural Information Processing Systems 19, 801 (2007)
5. Olshausen, B.A.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
6. Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 111–118 (2010)

7. Boureau, Y.L., Bach, F., LeCun, Y., et al.: Learning mid-level features for recognition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2559–2566. IEEE (2010)
8. Yu, K., Zhang, T., Gong, Y.: Nonlinear Learning using Local Coordinate Coding. In: NIPS, vol. 9, p. 1 (2009)
9. Wang, J., Yang, J., Yu, K., et al.: Locality-constrained linear coding for image classification. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3360–3367. IEEE (2010)
10. Lin, Y., Zhang, T., Zhu, S., et al.: Deep Coding Network. In: NIPS, pp. 1405–1413 (2010)
11. Lee, H., Ekanadham, C., Ng, A.Y.: Sparse deep belief net model for visual area V2. In: NIPS, vol. 7, pp. 873–880 (2007)

# Video Smoke Detection Based on the Optical Properties

Yingjing Wu<sup>\*</sup> and Ying Hu

Automation Research Centre, Dalian Maritime University, China

Wuyingjing\_work@163.com

hawk\_huy@sina.com

**Abstract.** Video smoke detection has many advantages such as high response speed and non-contact detecting. But the current video detection methods are either complicated or less reliable. A suitable method for ordinary video smoke detection by analyzing optical properties of smoky images is presented in this paper. The factors of optical properties such as scene radiance, medium transmission, path-length and total scattering coefficient were studied. Different scene radiances represent different objects. Using scene radiance helps us to recognize the suspected area that almost doesn't change which may include those smoky areas. What's more, it is found that the total scattering coefficient would increase along with the growing number of particles in the atmosphere caused by smoke, and lead the medium transmission to decrease. The decision rule based on this finding aims to narrow down the suspected smoky region. The experiment results show that this method is effective and practical.

**Keywords:** smoking detection, optical properties, dark channel prior, video image.

## 1 Introduction

Fire, bringing us the loss of many lives and economical damage, is a common problem faced by people around the world. Though the consciousness to prevent fire is being continuously raised, there is little chance to avoid all the disasters caused by fire. The only way to reduce losses is to detect and put out fire as soon as possible.

Conventional smoke detection using sensors monitors the changes of temperature and gas composition to judge fire. But it has its practical limits. To solve the problem like air flow, space span and judgment based on single justification, many scholars are studying on fire alarming based on video. At the beginning of burning, though produces no obvious flames, it would come up with some physical phenomena such as smoking and smouldering, so video smoke detection is able to find signals of fire and realize early alarming.

Many works have been done on video smoke detection. One strategy is to use colour and motion information to detect smoke from digital images. Chen and Huang et al. [1] judge smoke-pixel by chromaticity-based static and diffusion-based dynamic

---

<sup>\*</sup> Corresponding author.

characteristic decision rule. Another one is based on the characteristics of time and space. Toreyin et al. [2] propose a background subtraction, temporal and spatial wavelet transformation based smoke detection method. The area of decreased high frequency energy component is identified as smoke using wavelet transforms. Motion direction based method is a third one. A sliding time window is used in the work of Yuan [3] to generate a time sequence of motion orientation for each block. A 3D feature is extracted from the accumulation and main motion orientation computed according to the sequence. A Bayesian classifier is used for smoke detection. Wang Liu and Xie [4] analyze flutter features of the motion region extracted over a sliding time window, and then the neuron-fuzzy inference system is used to detect smoke, in which fuzzy rules and membership functions are trained according to the valid sample set. These methods still have their limits. (1) Since the colour of smoke is widely varied, it's a big challenge to figure out the substances that have the similar colour with smoke. A misjudgment would also be made even the method has considered both colour and motion information. For example, telling a white plastic bag flowing from far to near apart smoke, which only depend on the colour and the change of size, is far from easy. (2) Using wavelength analysis detection algorithm requires images have blur edges and translucent characteristics. Though in a certain environment its test result is mostly accurate, the algorithm cannot exert its effect in conditions that don't meet requirements. (3) Using the ratio of the circumference area and the change of the area as the characteristics of algorithms is more prone to misjudge when there are many moving objects in the scene. Chengjiang Long et al. [5] use transmission to detect smoke region by looking for the best threshold and obtain detailed information about the distribution of smoke thickness through mapping transmissions of the smoke region into a gray image. But the threshold value is hard to determine and the translucent smoke cannot be figured out if the threshold value is not proper.

In this page, a method based on optical properties via video is proposed to detect smoke. Optical thickness depends on two main factors - scattering coefficient and the wavelength of light. We find that in most of the video where the light changes smoothly, the raising scattering coefficient caused by a large increment in the particles of smoky region will finally reflects on the decreasing of the medium transmission obtained from smoky frame. Using this finding, the smokeless regions identified by the difference between the scene radiance of smoky frame and smokeless ones can be eventually narrowed down. The method has low computing complexity, low false alarming rate and high accuracy.

## 2 Backgrounds

### 2.1 Attenuation

The attenuation of a beam of light when it travels through the atmosphere causes the radiance of a scene point to fall as its depth from observer increases. Here we will introduce the derivation of the attenuation model given by Bouguer's exponential law [5]:

$$J(\lambda, x) = J(x)e^{-\beta(\lambda)d} \quad (1)$$

Sometimes, attenuation can also be expressed in terms of optical thickness as follows:

$$T(x) = -\beta(\lambda)d(x) \quad (2)$$

Where  $J(\lambda, x)$  is an attenuated irradiance describing scene radiance and its decay in the media,  $J(x)$  is the scene radiance. Different scene radiances correspond to different subjects, which is used as a rule for us to exclude some interference. Smoke has impact on the number of the particle in air, but it doesn't turn the background into other objects, so the radiance shouldn't change.  $d(x)$  is a path-length in  $x$  point,  $\beta(\lambda)$  is a total scattering coefficient and  $\lambda$  is a wavelength of light in  $x$  point. The optical thickness  $T(x)$  represents the ability of a light-wave with given wavelength to scatter in all direction. It is independence of distance in where the atmosphere is homogeneous. We find there's a relationship between the optical thickness and the smoky pixel. Since the total scattering coefficient  $\beta(\lambda)$  and the distance  $d$  cannot be directly measured from the image, we use the optical model and dark channel prior mentioned in Section 2.2 and 2.3 to figure out the medium transmission to represent the optical thickness. Thus we can use the medium transmission to judge smoky region.

## 2.2 Optical Model

According to the atmosphere physics, the atmosphere scatters light energy radiating from scene. The following optical model is widely used to describe the formation of a haze image [6, 7, 8, and 9]:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (3)$$

Where  $I$  is an observed intensity,  $t(x)$  is the medium transmission describing the portion of the light that not reaches the camera.  $A$  is the global atmospheric light which is set to be constant. Noted that,  $I$ ,  $J$  and  $A$  are all vectors in RGB color space. The first term  $J(x)t(x)$  on the right hand side of equation (3) is the attenuated irradiance which we have mentioned in Section 2.1 Equation (1). The second term  $A(1 - t(x))$  is called air-light resulting from the shift of the scene color lead by previously scattered light. In the Section 2.1, we have introduced attenuation model. Compare the first term  $J(x)t(x)$  in Equation (3) with Equation (1), we find:

$$t(x) = e^{-\beta(\lambda)d} = e^{T(x)} \quad (4)$$

Equation (4) clearly explains the reason we use medium transmission to represent optical thickness. We can get the value of the observed intensity and the global atmospheric light  $A$ -a constant value easily. As long as we obtain the scene radiance  $J(x)$ , can we calculate the medium transmission. In [10], Kaiming He illustrated that some of the approaches to obtain  $J(x)$  and  $t(x)$  by using Equation (3) are not very proper, and proposed a method to estimate  $t(x)$  which would be introduced in Section 2.3.

### 2.3 Dark Channel Prior

Most local patches in haze-free outdoor images contain some pixels which have very low intensities in at least one color channel. For an image  $J$ , we define

$$J_{dark}(x) = \min_{c \in \{r, g, b\}} (\min_{y \in \Omega(x)} (J^c(y))) \quad (5)$$

Where  $c$  is a color channel of  $J$  and  $\Omega(x)$  is a local patch centered at  $x$ . We call  $J_{dark}$  the dark channel of  $J$ . We assume that the transmission in a local patch  $\Omega(x)$  is constant and then take the min operation performed on three color channels independently in the local patch on Equation (3). Change the terms and we have:

$$t(x) = 1 - \frac{J_{dark}(x)}{A} \quad (6)$$

Kaiming He's work aims to remove haze. A transmission map shows the positions of each object in the image and medium transmission plays an important role in a soft matting algorithm. In this paper, we focus our attention on combining the transmission with the smoky area, in another word, to figure out their relationship. So finding out the relationship between smoky pixel and medium transmission is next step. Total scattering coefficient

We assume that the path-length is constant to a series of images. When there is a smoke, the change of the light decay in the smoking regions mainly depend on  $\beta(\lambda)$ . Thus, the model of the total scattering coefficient would be induced here. The wavelength has impact on the scattering of the atmosphere particle. So scattering coefficient has the following relationship with wavelength [8]:

$$\beta(\lambda) \propto \frac{1}{\lambda^\gamma} \quad (7)$$

We restrict that the light in the model changes smoothly, which can be satisfied under the smoky situation, then, the coefficient  $\beta$  is relevant to  $\gamma$ . The value of  $\gamma$  depends on the number of the particle in the atmosphere. Normally,  $0 \leq \gamma \leq 4$ . In pure air, the radius of particle is much smaller than the wavelength of light, so the scattering is weak. In this case,  $\gamma = 4$ ; While in the condition of heavy smoke, since radius of smoke particle is bigger than wavelength of light, the change of scattering coefficient lead by wavelength of light in the visible range could be very small, thus,  $\gamma \approx 0$ .

## 3 Smoke Obscuring Model

Smoke normally keeps other objects out in the form of opaque or translucent. But no matter how it looks like, smoke has no direct effect on the scene radiance  $J(x)$  and the air-light  $A$ . Only if there is a smoke does the scattering coefficient change, which means the transmission image would be different. A and B represent two frames of a video where records the whole process of a fire or the motion of some obstacles that

are similar to smoke. A, the one has no smoke in it, is called a smokeless image, while B is the one that we call smoky image no matter whether it records real smoke or some obstacles. Their observed intensities are shown in equation (8) and (9). The following two calculations are all operated in RGB colour space separately.

$$I_A(x) = J_A(x)t_A(x) + A(1-t_A(x)) \quad (8)$$

$$I_B(x) = J_B(x)t_B(x) + A(1-t_B(x)) \quad (9)$$

The scene radiances of the smokeless and the smoky images are equivalent theoretically, which means the two images have nothing different except for the appearing of smoke or interference objects. Using Equation (9) minus Equation (8) to get the difference between the observed intensity, we obtain:

$$\Delta I(x) = (J(x) - A)(t_B(x) - t_A(x)) \quad (10)$$

We only focus on the medium transmission  $t(x)$ , for its change is the main factor that causes the change of  $\Delta I(x)$ . We can directly figure out the medium transmission by Equation (6). But our purpose is to try to link optical thickness with smoky pixel by medium transmission. On the left side of the equation is the value of the medium transmission difference obtained through dark prior theory, while on the right side is the medium transmission difference in forms of Equation (4):

$$\Delta t(x) = t_B(x) - t_A(x) = e^{-\beta_B(\lambda)d_B(x)} - e^{-\beta_A(\lambda)d_A(x)} \quad (11)$$

If there is some smoke, which means the path-length  $d$  of these series image is changeless, the scattering coefficient would be bigger which is known in Section 1.4. So we have:

$$\Delta t(x) = e^{-\beta_B(\lambda)d(x)} - e^{-\beta_A(\lambda)d(x)} < 0 \quad (12)$$

If the path-length doesn't change, neither dose the scattering coefficient, we can get:

$$\Delta t(x) = e^{-\beta_B d_B(x)} - e^{-\beta_A d_A(x)} = 0 \quad (13)$$

Equation (13) illustrates the scene in image A and B at the  $x$  point are the same.

And if the scene radiance could not identify the distracters, then we will neglect the changes of the scattering coefficient caused by the number of the particle in the air and only focus on the distance change. The path-length  $d$  becomes smaller for the blocking of the obstacles. We can have:

$$\Delta t(x) = e^{-\beta(\lambda)d_B(x)} - e^{-\beta(\lambda)d_A(x)} > 0 \quad (14)$$

Through the above equations, we can determine whether each pixel in the video is the region blocked by smoke by  $D(x)$ .

$$D(x) = \begin{cases} 1 & \Delta t(x) < 0 \\ 0 & \Delta t(x) \geq 0 \end{cases} \quad (15)$$

In practice, the distance between scene and observer has little chance to become larger. In the wild wind, the leaves of branches or other object movement interference can be ruled out through the comparison of the scene radiance. In cities, car lights and other lights may interfere test results. For example the car lights will reduce the transmission which should be recognized as smoky pixels by Equation (15). But before we use Equation (15), a preliminary region of smoke has already been decided by scene radiance which rules out the car lights because the strong contrast between the lights and the original black background.

## 4 Method for Detection

### 4.1 Judgment of Suspicious Area

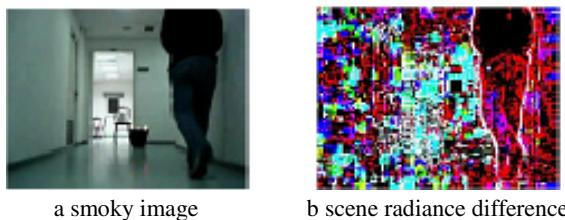
Compare each pixel of the smokeless image and the smoky image to extract the suspicious areas, the concrete process is:

Get the medium transmission of the each image. Substitute the medium transmission into Equation (3),

We obtain:

$$J(x) = \frac{I(x) - I_b}{t(x)} \quad (16)$$

- (1) Get their scene radiance respectively.
- (2) Subtract the scene radiance to get the difference.
- (3) Since the scene radiance is only related to the object in the scene itself and has nothing to do with path-length and air-light, whether the objects are the same is merely depended on whether the scene radiance is equivalent. Eliminate the areas of which the scene radiance are different, and then keep the rest as the suspicious regions. The dark parts in Fig. 1b are the suspicious regions.

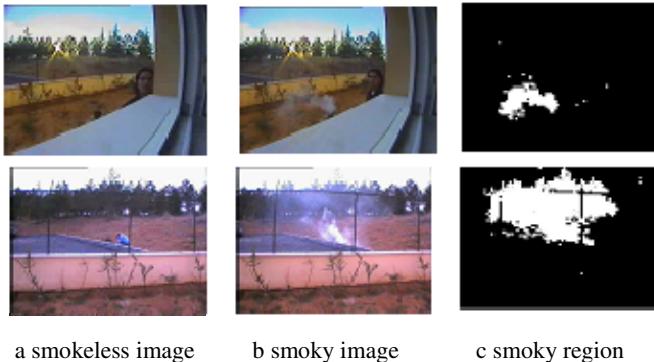


**Fig. 1.** Using scene radiance to obtain suspicious region

## 4.2 Extract the Properties of Smoke and Judge the Smoky Region

Now a part of smokeless regions have been excluded, but there are still some portions haven't been eliminated. We need to use the medium transmissions of these suspicious regions and the smoke obscuring model to select the smoky region. The specific processes are:

- (1) Calculate the difference between the medium transmissions.
- (2) Use Equation (15) to decide whether the area has smoke.



**Fig. 2.** Using optical properties to declare the smoky region smokeless image

## 5 Experimental Results

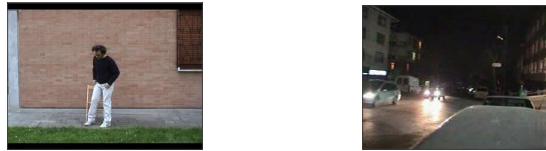
To verify the robustness and effectiveness of the video smoke detection algorithm, we test multiple  $320 \times 240$  videos. The optional videos cover a variety of experimental scenarios and different sources of ignition. The algorithm is tested on windows 7 operating system, using matlab as a software platform. Some of the testing images are coming from <http://imagelab.ing.unimore.it/visor>, <http://signal.ee.bikent.edu.tr/VisiFire/Demo/SmokeClips/> and the other come from the video made by ourselves. In order to judge smoke and have a more visible feeling on the results, white rectangles are used to note if there's a smoke in the frame by selecting regions where there are more than a certain number (threshold) of white pixels in the image after using the smoke obscuring model.

Due to space limitations, only 7 groups of video smoke detection results are listed. The statistical data come from 9 sets of smoky test video and six sets of smokeless test video. Test results are shown in Figure 3 and Figure 4. Figure 4 is a result under the condition of light interference which explains the algorithm can effectively suppress the non-smoke interference and reduce false alarming rate.



a sWastesmoke

b Jfz

**Fig. 3.** The smoke detection results using our algorithm

a Carlights1

b Man with dog

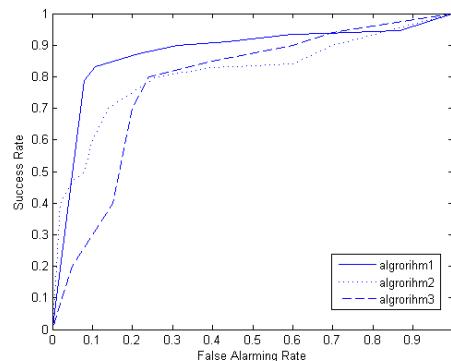
**Fig. 4.** The smoke detection results with disturbed by different objects**Table 1.** The test video data

Name of video	frames	description of video
sWastesmoke	2160	smoky,close view, heavy smoke
sWindows	1384	smoky,outdoor, strong light
sEmptyR1	1448	smoky,indoor, strong light
Jfz	3000	smoky,indoor, problem with raytracing
Carlights1	1450	smokeless,nighttime, traffic, strong light
Carlights2	1352	smokeless, nighttime, traffic, strong light
Man with dog	1525	smokeless,outdoor, moving objects

In order to assess our algorithm, ROC (receiver operating characteristics) curve is used as an index of evaluation. ROC curve describes the relationship between the success rate and the false alarming rate. The false alarming rate is represented by the abscissa and success rate the ordinate. The success rate and false alarming rate change along with the threshold. Connecting points with smooth curve makes ROC curves. The area under the ROC curve (AUC. Area under the curve) can value the performance of the algorithm in the smoke detection. The closer AUC value is to 1, the better performance of the algorithm will be.

To compare, we implement the algorithm based on colour feature and the work [3] based on the accumulation and the main motion orientation by extracting 3d feature after using naive Bayes Classifier for smoke detection. They are respectively denoted as algorithm 2 and algorithm 3. The algorithm of this paper is labeled as algorithm 1. Three kinds of algorithms are on the same data set for feature extraction, training and testing. The ROC curve is shown in figure 5.

AUC of the algorithm 1 is 0.889, the algorithm 2 0.786 and the algorithm 3 0.827. And overall, the curve of algorithm 1 is above algorithm 2 and 3. So our algorithm can maintain a relatively low false alarm rate and, meanwhile, keep an ideal success rate.



**Fig. 5.** The comparison among the ROC curves of different algorithms

## 6 Conclusion

This paper proposes the video smoke detection based on the optical properties. The preliminary model to obtain the suspected areas is constructed by an optical model combined with the dark colors theory. We select the rest regions with the information of optical thickness, compare the physical characteristics of smoke with other interferences and effectively identify smoke areas. This method is proved to have the simplicity of computational efficiency and relatively low false alarming rate than other current video smoke detection methods. There is underreporting phenomenon when the proposed method is used to detect in a white background. To extract more representative smoke feature model and to further improve the accuracy of the smoke detector is the focus of our future work.

## References

1. Chen, T.H., Yin, Y.H., Huang, S.F., et al.: The Smoke Detection for Early Fire-Alarm System Based on Video Processing. In: Proceeding of 2006 Internet Conference on Intelligent Information Hiding and Multimedia Signal Processing, USA, pp. 427–430 (2006)
2. Toreyin, B.U., Dedeoglu, Y., Cetin, A.E.: Wavelet-Based Real-Time Smoke Detection in Video. In: Proceeding of 13th European Signal Processing Conference, Piscataway, pp. 4–8 (2005)
3. Yuan, F.-N., Zhang, Y.-M., Liu, S.-X., et al.: Video Smoke Detection Based on Accumulation and Main Motion Orientation. Journal of Image and Graphics 13(4), 808–813 (2008)

4. Wang, T., Liu, Y., Xie, Z.-P.: Flutter Analysis Based Video Smoke Detection. *Journal of Electronics and Information Technology* 33(5), 1024–1029 (2011)
5. Long, C., Zhao, J., Han, S., Xiong, L., Yuan, Z., Huang, J., Gao, W.: Transmission: A New Feature for Computer Vision Based Smoke Detection. In: Wang, F.L., Deng, H., Gao, Y., Lei, J. (eds.) *AICI 2010. LNCS (LNAI)*, vol. 6319, pp. 389–396. Springer, Heidelberg (2010)
6. Narasimhan, S.G.: Models and Algorithms for Vision through the atmosphere. In *Columbia Univ. Dissertation* (2004)
7. Narasimhan, S.G.: Interactive Deweathering of an Image Using Physical Models. In: *ICCV Workshop on Color and Photometric Method in Computer Vision*. IEEE Computer Society (2003)
8. Shuai, F., Yong, W., Yang, C., et al.: Restoration of Image Degraded by Haze. *Acta Electronica Sinica* (10), 2279–2284 (2010)
9. Cheng, G., Wang, T., Zhou, H.-Q.: A Novel Physics-based Method for Restoration of Foggy Day Images. *Journal of Image and Graphics* 13(5), 888–893 (2008)
10. Fattal, R.: Single image dehazing. *SIGGRAPH2008*, LosAngeles: ACM Transactions on Graphics 27(3), 1–9 (2008)
11. He, K., Sun, J., Tang, X.: Single Image Haze Removal Using Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12), 2341–2353 (2011)

# Discovery of the Topical Object in Commercial Video: A Sparse Coding Method

Yunhui Liu, Huaping Liu, and Fuchun Sun

Department of Computer Science and Technology, Tsinghua University, Beijing, China  
State Key Laboratory of Intelligent Technology and Systems, Beijing, China

**Abstract.** In this paper, we propose a topical object discovery method in commercial video. This method utilizes the objectness measure to generate the object candidates from the key-frames of the video. Then a sparse coding method is developed to discover the most topical object. Such a method can provide ranked results and therefore we can easily select the most topical object. The experimental validation on 10 videos shows that the sparse coding method performs better than existing topic mining methods.

**Keywords:** Topical Object Discovery, Objectness, Sparse Coding.

## 1 Introduction

With the popularization of Internet and TV, people are now meeting more and more commercial video clips than ever before. These large collections of commercial video clips are intrinsically difficult to summarization, due to their diversified contents. Automatic object discovery, which locate the topical object occurs frequently in such video, becomes very important to help people to [1] rapidly grasp the main content of the video [2] and evaluate the quality of the commercial video. Such a technology can also be used to summarize the video clips. The main difficulty in automatic object discovery in video is how to effectively extract certain objects of the video while preserving the essential content of the original video.

An early work to discover the object in image collections is [3]. In that work the authors segmented each image at different scales and assume that the topical object lies in some segments. Then the Latent Dirichlet Allocation(LDA) is used to discover the topics. Finally, the segments which are the most similar to the discovered topics are extracted to represents the topical objects. In [4], such a method was extended to short videos and a modified LDA method was developed. Both the work share a common demerit. i.e., the number of the topics should be prescribed by the designer. In [5], a different Non-negative Matrix Factorization method is proposed to solve similar problem. Also, the rank of the matrix is a parameter which should be determined by the users. In addition, all of the work depends on the image segments, which represents a very coarse representation for the object. Therefore, the extract results are difficult to evaluate. The sub-graph mining [6] and the method in [2] were not depends on the segments. But such method cannot extract complete object region. In [1], a survey about this problem was given.

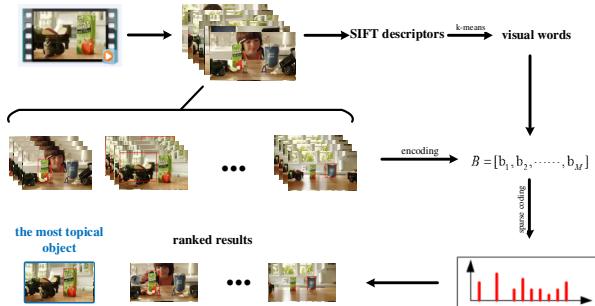
In this paper, we resort to the objectness which was originated by [7] to develop more effective object discovery tools. Since the objectness is introduced, the discover topic will be more close to the true object. The main contributions of this work are listed as follows:

- (1) The objectness is introduced to discover object candidates in the video key-frames. To the best knowledge of the authors, there is no related report on such applications of objectness.
- (2) A sparse coding model is to discover the most topical objects. A great disadvantage of such a method is that there is no prescribed number of topics should be given.
- (3) We collect more than 10 commercial video clips from the YouTube and made an extensive evaluation of the proposed method, and the state-of-the-art.

The rest of this paper is organized as follows: Section 2 provides the overview of proposed method; Section 3 describes the sparse coding problem formulation; Section 4 presents the experimental results and Section 5 gives conclusions.

## 2 Object Candidates Generation

Fig. 1 shows the work flow of the topical object discovery procedure. After getting the key-frame from the video, we represent images using affine covariant regions, described by SIFT [8] and quantized into 1000 visual words. On the other hand, for each key-frame, we generate object candidates using the work in [7] to get a high chance of obtaining good bounding boxes that will contain the interested objects. Once the visual words are computed for an image, each of object candidates is represented by a Bag-of-Words (BoW) histogram of visual words contained within the bounding box. Then the sparse coding method comes into play. It ranks the object candidates and then we can select the most significant candidates according to the coding results.



**Fig. 1.** The work flow of proposed approach

According to [7], the object candidate generation method starts with finding an informative prior that captures the potential salient regions from images. This method focuses on algorithms that are able to handle general object appearances without category specific information. To this end, the objectness algorithm is developed to find a set of object candidates in input images.

Specifically, the objectness algorithm finds a set of object candidates represented by bounding boxes, together with the confidence scores, for each input image. It adopts four different low-level cues to learn if a certain bounding box contains an object or not, which we give a brief explanation to the cues adopted in the method as follows for completeness. For more details, please refer to [7]. Concretely speaking, the multiscale saliency cue utilizes the spectral residual of the Fourier Transform on multiple scales to find regions with unique appearances within the image. The color contrast cue computes the dissimilarity of the color distribution of a candidate bounding box with that of its surrounding area. The edge density cue computes the density of the edges (computed by Canny edge detection) near the borders of the candidate bounding box. Finally, the superpixel straddling cue computes the agreement between the candidate bounding box and the superpixels obtained by [9]. Since pixels in the same superpixel often belong to the same semantic group (either the object or the background), for a good object candidate most superpixels should lie mostly either inside or outside the bounding box, and should not cross the boundary.

In Fig. 2 we give an example by using such a method. The algorithm is able to capture all the correct location of salient objects in the key-frame with 20 sampled windows. For robustness, we filter the too big ones which are too coarse for actual objects and small ones which generally contain no complete objects. Some unsatisfactory results are illustrated in Fig. 3. The objectness work [7] tries to find all objects in image, so inevitably some too big or small boxes are generated. However, they can be safely filtered out by using some heuristic rules according the size or location prior.



**Fig. 2.** Exemplar results of bounding boxes. We sample 20 windows and show some good examples separately on the right side.



**Fig. 3.** Some bad bounding boxes. Too small ones (the left 2) generally contain no complete objects and too big ones (the right 2) are coarse for actual objects.

### 3 Sparse Coding Method

After we get all of the object candidates, the remaining thing is to extract the most topical object from the candidate set. The topical objects can be regarded as a small set consists of a collection of representative objects selected from the underlying candidate set.

Therefore, topical objects selection is equivalent to how to select an optimal subset from the entire candidate set under certain constraints. Consider a matrix

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N] \in R^{d \times N}, \quad (1)$$

where each column vector denotes a feature vector for one candidate window. The task is to find an optimal subset

$$\bar{\mathbf{B}} = [\mathbf{b}_{k_1}, \mathbf{b}_{k_2}, \dots, \mathbf{b}_{k_n}] \in R^{d \times n}, \quad (2)$$

where  $k_1, k_2, \dots, k_n \in \{1, 2, \dots, N\}$ , to satisfy some optimization performance. That is to say, the topical objects selection problem is equivalent to solving the following optimization problem

$$\min_{\bar{\mathbf{B}}} f(\mathbf{B}, \bar{\mathbf{B}}), \quad (3)$$

where  $f(\cdot, \cdot)$  is a prescribed optimization objective.

The selected objects should be representatives of all of the object candidates, i.e., a sample  $\mathbf{b}_i$  should be represented as a linear combination of  $\mathbf{b}_{k_1}, \mathbf{b}_{k_2}, \dots, \mathbf{b}_{k_n}$ . To characterize this reconstruction capability, the following cost should be minimized to use such topical objects to reconstruct the whole candidate set:

$$\min_{\bar{\mathbf{B}}, \bar{\mathbf{X}}} \frac{1}{2} \|\mathbf{B} - \bar{\mathbf{B}}\bar{\mathbf{X}}\|_F^2, \quad (4)$$

where  $\|\cdot\|_F$  is the Foubenious norm of the matrix, and  $\bar{\mathbf{X}} \in R^{n \times N}$  is the coefficient matrix.

The problem is that the coefficients  $\bar{\mathbf{X}}$ , as well as the index set  $\{k_1, k_2, \dots, k_n\}$ , are unknown. Hence, one starts by using all columns of  $\mathbf{B}$  to describe  $\mathbf{B}$  itself, i.e.,  $\mathbf{X} \in R^{N \times N}$  should be found to satisfy

$$\mathbf{B} = \mathbf{B}\mathbf{X}. \quad (5)$$

In addition, the number of the selected topical objects should be as small as possible. To tackle this problem, a straightforward approach is to minimize the following objective function

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{B} - \mathbf{B}\mathbf{X}\|_F^2 + \lambda \|\mathbf{X}\|_{2,0}, \quad (6)$$

where  $\lambda$  is used to balance different penalty terms. The first term  $\|\mathbf{B} - \mathbf{B}\mathbf{X}\|_F^2$  is used to evaluate the reconstruction error, and in the absence of the second term, unconstrained optimization will lead to the trivial solution  $\mathbf{I}$ . The second term  $\|\mathbf{X}\|_{2,0}$  is defined as

$$\|\mathbf{X}\|_{2,0} = \sum_{i=1}^N \delta(\|\mathbf{X}^{(i)}\|_2), \quad (7)$$

where  $\delta(\cdot)$  is the Dirac operator and  $\mathbf{X}^{(i)}$  represents the  $i$ -th row of  $\mathbf{X}$ . This definition indicates that  $\|\mathbf{X}\|_{2,0}$  counts the number of nonzero rows of  $\mathbf{X}$ . By adding the  $L_{2,0}$  norm term into the objective function, the trivial solution can be avoided and the obtained solution will be row sparse, i.e., most of its rows are zero vectors. However, such an approach is of little practical use, since the optimization problem is NP-hard as its solution requires a combinatorial search which grows faster than polynomial as the dimension  $N$  grows. A natural alternative is to use the  $L_{2,1}$  norm to replace the  $L_{2,0}$  norm [10], resulting the following convex optimization problem [11]:

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{B} - \mathbf{BX}\|_F^2 + \lambda \|\mathbf{X}\|_{2,1}, \quad (8)$$

where

$$\|\mathbf{X}\|_{2,1} = \sum_{i=1}^N \|\mathbf{X}^{(i)}\|_2 \quad (9)$$

is the sum of the  $L_2$  norm of the rows of  $\mathbf{X}$ . The optimization problem in Eq.(8) can be efficiently solved by using gradient methods [12] or alternating direction method of multipliers [13]. After obtaining the value of  $\mathbf{X}$ ,  $\|\mathbf{X}^{(i)}\|_2$  is used to evaluate the possibility that the  $i$ -th candidate box acts as the topical object. If  $\mathbf{X}$  is indeed row sparse, i.e., most of the rows in  $\mathbf{X}$  are zero vectors, then one simply selects the candidates which corresponds to the non-zero row in  $\mathbf{X}$  to be the topical objects.

The whole procedure can be summarized as follows: First, the feature vectors are extracted from the object candidates to form the matrix  $\mathbf{B}$ . Then the sparse coding is performed to obtain the coefficient matrix  $\mathbf{X}$  and the resulting weight  $w_i = \|\mathbf{X}^{(i)}\|_2$  for the  $i$ -th candidates. Upon the weight curve one can get the ranked results  $\{k_1, k_2, \dots, k_N\}$ , where  $w_{k_1} \geq w_{k_2} \geq \dots \geq w_{k_N}$ . The first several object candidates of which index number are  $k_1, k_2, \dots, k_n$  are provided as the selected topical objects.

The objective function in Eq.(8) is essentially convex and therefore we can effectively solve it. According to [10], it can be solved by minimizing the following objective function iteratively:

$$\frac{1}{2} \|\mathbf{B} - \mathbf{BX}\|_F^2 + \lambda Tr(\mathbf{X}^T \mathbf{DX}), \quad (10)$$

where  $Tr(\cdot)$  is the trace of a matrix,  $\mathbf{D}$  is a diagonal matrix with the  $i$ -th diagonal element as

$$d_{ii} = \frac{1}{2 \|\mathbf{X}^{(i)}\|_2}, \quad (11)$$

Note that in practice,  $\|\mathbf{X}^{(i)}\|_2$  could be very close to zero. In this case, one can follow the traditional regularization way and define the diagonal elements of  $\mathbf{D}$  as

$$\frac{1}{2 \|\mathbf{X}^{(i)}\|_2 + \varsigma}, \text{ where } \varsigma \text{ is a small constant.}$$

Setting the derivative of Eq.(10) to zero leads to

$$(\mathbf{B}^T \mathbf{B} + \lambda \mathbf{D}) \mathbf{X} = \mathbf{B}^T \mathbf{B}, \quad (12)$$

and therefore  $\mathbf{X}$  can be represented as

$$\mathbf{X} = (\mathbf{B}^T \mathbf{B} + \lambda \mathbf{D})^{-1} \mathbf{B}^T \mathbf{B}. \quad (13)$$

Unfortunately, since  $\mathbf{D}$  is dependent of  $\mathbf{X}$ , Eq.(13) does not give closed-form solution to  $\mathbf{X}$ .

To tackle this problem, an iterative algorithm is proposed in Algorithm 1. The termination is declared when there are only negligible changes in the objective function value. Empirically, the proposed iterative algorithm works well.

---

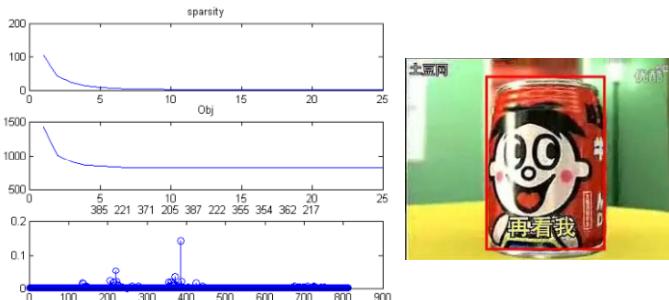
**Algorithm 1.**


---

**Input:** Data set  $\mathbf{B} \in R^{d \times N}$

**Output:** Solution  $\mathbf{X} \in R^{N \times N}$

- 1: Set  $t = 0$ . Initialize  $\mathbf{D}_t \in R^{N \times N}$ , as identity matrices.
  - 2: **While** Not convergent **do**
  - 3:     Calculate  $\mathbf{X}_{t+1} = (\mathbf{B}^T \mathbf{B} + \lambda \mathbf{D}_t)^{-1} \mathbf{B}^T \mathbf{B}$ .
  - 4:     Calculate the diagonal matrix  $\mathbf{D}_{t+1}$  of which the  $i$ -th diagonal elements are  $\frac{1}{2 \|\mathbf{X}_{t+1}^{(i)}\|_2}$ , respectively.
  - 5:     Set  $t = t + 1$ .
  - 6: **end while**
- 



**Fig. 4.** The sparse coding results on the video *Wangwang*. On the left panel, the first row shows the sparsity and the objective function value versus the iteration number, respectively. The last row shows the coding coefficient  $\|\mathbf{X}^{(i)}\|_2$  versus the candidate index  $i$ . The right panel shows the selected topical object box (see the red box) which is indexed by 385.

After deriving  $\mathbf{X}$ , one can use  $\|\mathbf{X}^{(i)}\|_2$  to rank the object candidates. The larger  $\|\mathbf{X}^{(i)}\|_2$  is, the more important this object candidate is. The user can either select a fixed number of the most important candidates or set a threshold and select the candidates whose  $\|\mathbf{X}^{(i)}\|_2$  is larger than this value. In this work, we are interested the most topical object and therefore we select the object candidate with the largest confidence. Fig. 4 shows iteration details of proposed method on a representative video *Wangwang*. From this figure we find that the iteration procedure converges well.

## 4 Experimental Results

### 4.1 Video Datasets

To evaluate our approach, we download 10 commercial videos from youtube.com and youku.com and try to find the most topical object one by one. For each video sequence, we set ground truth object manually. The information about the investigated sequences is shown in Table 1.

**Table 1.** The information about the investigated sequences

No.	Sequence Name	Dur. (s)	Resolution	fps
1	Banana Milk	30	512×288	15
2	Daliyuan	29	416×328	15
3	Dominos	24	456×360	25
4	Lanyueliang	32	448×336	15
5	MM	90	512×288	15
6	Nestle	28	512×288	15
7	Sultana Bran	28	512×288	15
8	Vemma Nutrition	23	640×360	23
9	Wangwang	40	320×240	15
10	Savia	30	640×360	25

### 4.2 Experimental Setting

To obtain the object representation for video, we first sample key-frames from each video at one for every ten frames. SIFT features are extracted from each key-frame as [3]. For each sequence, the local features are quantized into  $V = 1000$  visual words by the  $k$ -means clustering. Then for each key-frame, we generate 20 image windows using the source code of the objectness measure [7] and choose the sub-images in the window whose area is more than 1/50 or less than 1/2 of the entire image as object candidates. After then each candidate is described by the BoW representation.

Once getting the most topical object, for each key-frame that contains the ground truth bounding box, we compare the K-L divergence of the topical object and candidate boxes and select the candidate who achieves minimum distance.

To quantify the performance, we manually label the ground-truth bounding boxes of the topic objects in each image. Let  $DR$  and  $GT$  be the discovered object and the bounding boxes of ground truth, respectively. The performance is measured by

$$score = \frac{|GT \cap DR|}{|GT \cup DR|}, \quad (14)$$

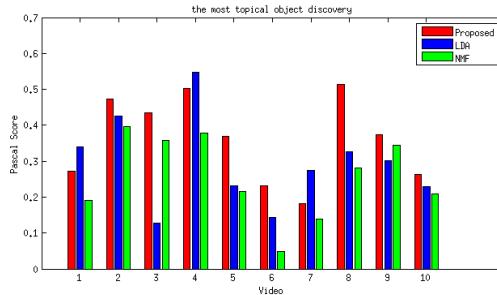
which measures the accuracy of topic object discovery. To calculate the score for one video, the score is first calculated for each key frame which contains ground truth bounding box and then the average value of those key frames is used to evaluate the whole video.

### 4.3 Comparison with Other Approaches

We compare our video object discovery method with the other two methods:

1. LDA method. This method was firstly proposed in [14]. It uses unordered BoW representation to automatically discover topics. Following the approach of using clustered affine-invariant point descriptors as “visual words”, Ref.[3] uses LDA to discover objects in image collections. In this method, the number of topics should be prescribed. In [4], this parameter is empirically set to 8.
2. NMF method, which generates a non-negative representation of data through matrix decomposition, was used in [5] to find representative classes from a collection of un-annotated images. In this method, a key parameter which characterizes the dimension of matrix should be prescribed. In [6], this parameter is set to 35.

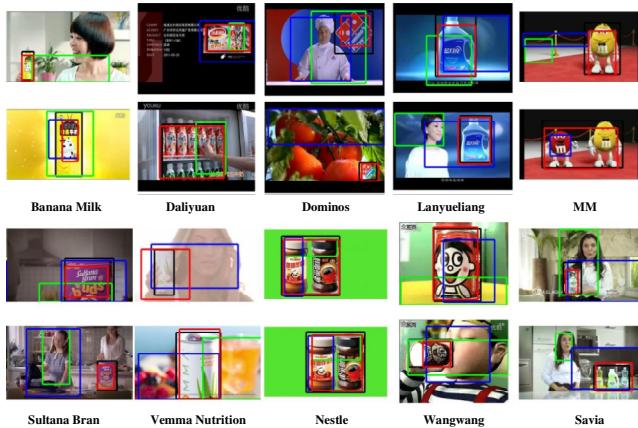
For a fair comparison, instead of multiple segments which was used in [3] and [5], we use object candidates obtained in Section 4.2 as the documents input. After discovering object topics using LDA following the work in [14], we select the topic which has the largest weight as the most topical object. The weight of the topic can be given by summing all documents’ topic distribution. The visual words and other settings are same as our method. For NMF approach [5], the evaluation procedure simply follows LDA.



**Fig. 5.** The performance comparison of three approaches

As shown in Fig. 5, our proposed approach outperforms both LDA approach and NMF approach in terms of the score defined in (14) for most topical object discovery, with an average score of 0.36 (Proposed) compared to 0.29 (LDA) and

0.26 (NMF), respectively. LDA approach has weak ability to generate the most topical object from discovered topics, although the largest weight method seems reasonable. NMF has the same problem. On the contrary, the proposed method sorts the object candidates during the global optimization and consider the information of bag-of-words more directly. It is interesting to note that the sparse coding method performs worse than LDA in certain videos. The main reason lies in that, in some cases, the topical object shows in different forms in video. However, in order to match with all the topical objects in different key-frames, our algorithm seems to find the most effective part on the topical object, which may occupies a small part of the ground truth bounding box.



**Fig. 6.** Sample results of evaluation. Black bounding boxes represent the ground truths. Red box is the result of proposed approach. Green and blue are for LDA and NMF approaches. Some green or blue boxes are overridden. However, the boxes rendering order is green, blue, red, that is to say the overridden result at most has the same performance as red box.

Fig. 6 shows sample results of most topical object discovery. For one video, we show discovered most topical objects on two key-frames. It can be seen that the most topical object discovered by the proposed approach can catch the ground truth object more accurately than other two approaches at different key-frames. For example, in video *Domino*, our algorithm captures the ground truth brand accurately while LDA and NMF's results seem coarse, only capturing a small part of brand.

## 5 Conclusions

This work utilizes the objectness to develop more effective object discovery tools. The objectness is introduced to discover object candidates in the video key-frames. Then a sparse coding model is to discover the most topical objects. A great disadvantage of such a method is that there is no prescribed number of topics should be given. Finally, we collect more than 10 commercial video clips from the YouTube and perform an extensive evaluation of various methods.

**Acknowledgements.** This work was supported in part by the National Key Project for Basic Research of China under Grant 2013CB329403; in part by the National Natural Science Foundation of China under Grants 91120011, 61210013; in part by the Tsinghua Self-innovation Project under Grant 20111081111; and in part by the Tsinghua University Initiative Scientific Research Program under Grant 20131089295.

## References

1. Wang, H., Zhao, G.: Visual pattern discovery in image and video data: a brief survey. *WIREs Data Mining Knowl. Discov.* 4, 24–37 (2014), doi:10.1002/widm.1110
2. Yuan, J., Zhao, G., Fu, Y., Li, Z., Katsaggelos, A.K., Wu, Y.: Discovering Thematic Objects in Image Collections and Videos. *IEEE Transactions on Image Processing* 21(4) (2012)
3. Russell, B. C., Efros, A. A., Sivic, J., Freeman, W.T., Zisserman, A.: Using Multiple Segmentation to Discover Objects and Their Extent in Image Collections. In Proc. of Computer Vision and Pattern Recognition (CVPR) (2006)
4. Zhao, G., Yuan, J., Hua G.: Topical Video Object Discovery from Key Frames by Modeling Word Co-occurrence Prior. In Proc. of Computer Vision and Pattern Recognition (CVPR) (2013)
5. Tang, J., Lewis, P.H.: Non-negative Matrix Factorisation for Object Class Discovery and Image Auto-annotation. In: Proc. of the 8th ACM International Conference on Image and Video Retrieval (CIVR)(2008)
6. Zhao, G., Yuan, J.: Discovering Thematic Patterns in Videos via Cohesive Sub-graph Mining. In: 11th IEEE International Conference on Data Mining (2011)
7. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the Objectness of Image Windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2012)
8. Lowe, D.: Object recognition from local scale-invariant features. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1150–1157 (1999)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. *IJCV* 59(2), 167–181 (2014)
10. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint l2,1 norm minimization. In: Proc. of Advances in Neural Information Processing Systems (NIPS), pp. 1–9 (2010)
11. Cong, Y., Yuan, J., Liu, J.: Abnormal Event Detection in Crowded Scenes using Sparse Representation. *Pattern Recognition* 46(7), 1851–1864 (2013)
12. Cong, Y., Yuan, J., Luo, J.: Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Trans. on Multimedia* 14(1), 66–75 (2012)
13. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In Proc. of Computer Vision and Pattern Recognition (CVPR), pp. 1600–1607 (2012)
14. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (2003)

# **Study the Moving Objects Extraction and Tracking Used the Moving Blobs Method in Fisheye Image**

Jianhui Wu<sup>\*</sup>, Guoyun Zhang, Shuai Yuan, Longyuan Guo, and Mengxia Tan

College of Information and Communication Engineering, Hunan Institute of Science and Technology, Yueyang, 414006, China  
wujhlf@foxmail.com

**Abstract.** This paper discusses a method of moving object detection and tracking in fisheye video sequences which based on the moving blob method. The fisheye lens has a very large angle of view and it has a better effective used at the no blind surveillance system, but the big distortion of the fisheye image that makes it difficult to achieve the intelligent function. In this paper some algorithms had been discussed which about detect and track the moving objects in the fisheye video sequences. It was divided three steps to discuss the processing algorithm. Firstly, the method of how to calculate the moving blob was discussed in fisheye image. This method included four main algorithms which are the background extracted algorithm, background updated algorithm, the algorithm of fisheye video sequence subtracted with the background to get the moving blobs, the algorithm of remove the shadow of blobs in RGB space. Secondly, the algorithm of how to determine every extracts blob are the real moving object was designed through calculated the pixels with threshold, it can discard the fault moving object. Lastly, the algorithm of tracking the moving objects was designed which based on the moving blobs of selected through calculated the geometry center of blobs. The experiment indicated that every algorithm have a better processing effective to the moving object in fisheye video sequences. The moving object can be detected effectively and stable. When too many objects are at the edge of the image, it is difficult to track every object because of the adhesions which are influenced by the large distortion. This method can be used in the large area fisheye surveillance system when there are not too many objects moving simultaneously.

**Keywords:** Moving blob, fisheye, object detection and tracking, algorithm.

## **1 Introduction**

The fisheye lens has an advantage in the large area of video surveillance because it has a large field of view. However, the circular fisheye image has a large distortion because of the special imaging principle of the fisheye lens. So the fisheye image processing is very difficult to achieve the ideal effect by using the normal digital image processing algorithm. In order to realize the processing of fisheye image effective, the main methods

---

<sup>\*</sup> Corresponding author.

are correction the fisheye image to 2D normal image, and then use the traditional image processing algorithm to deal with it. In the study of fisheye image correction, many researchers of domestic and foreign have carried out a lot of fruitful works, such as the fisheye image correction method based on the spherical perspective projection constraint[1], and established the transform model of the fisheye image to perspective projection image[2], as well as some researches of the fisheye image calibration and correction algorithms[3-5]. These research methods can be better to realize the calibration and transform the circular fisheye image to plane image. However, the corrected image will loss a lot of information especial at the edge of the fisheye image, and it need more time to run the correction algorithm. So it is not reach the no loss of information processing, and will bring great burden to the hardware system in practical application. When add the subsequent algorithm like object detection and others, the hardware system will difficult to processing real time. So in the fisheye video surveillance system, the method of corrected fisheye image firstly and further processing the 2D normal image use traditional algorithm will not achieve the "seamless" (because of information loss) monitoring and to realize the "real-time" processing of the fisheye image with optimal algorithms. Therefore, it is necessary to study the direct processing algorithms for the fisheye image, and designed a series of algorithms for direct fisheye image processing. In the area of traditional moving object detection, it has a lot of algorithms like Optical Flow method[6,7], Meanshift method and Camshift method[8-10] et al. These methods can detection the moving object from the normal video sequences, but it difficult to detect the moving object in fisheye video sequences. Especially, in order to enable accurate motion tracking for a large set of points in the video as close to real time as possible. To ensure accuracy, many methods only track a sparse set of points. In this paper, the improved moving object detection algorithm of fisheye sequences was proposed based on the moving blob method. Some moving blobs can synthesis like some pixel of foreground in the fisheye sequences which have the same motion attribute. Then we can detection and tracking these moving blobs, and realize the moving object detection and tracking in the fisheye images. The experiment indicates that it has a good detection effective about the moving foreground in fisheye video sequences and has a good practical value.

## 2 The Principle of Moving Blobs Method

### 2.1 Improved Algorithm of Mean Background Extraction

For the acquisition of fixed focal fisheye video sequences, the times of background more than the prospect. Therefore, we can sample in the video sequence with a period of time, such as 3 frames per second. For each pixel of the sample image frame, we calculated the averaged. According to the principle we know the average value will be closed to the background. Especial the background is more than the prospect, the average value is more closed to background. At the same time, the noise will be suppressed because of the average method. The specific algorithm is as follows:

- (1)  $N$  fisheye image frames  $F_i$  was got through sample, here  $i=1,2,\dots,N$ .
- (2) For every pixel  $(x,y)$  in every frames, calculated the background with:

$$B(x, y) = \frac{\sum_{i=1}^N F_i(x, y)}{N} \quad (1)$$

This simple method was depended on the sample frames, and it is not so good for many moving objects in video. If we can remove the pixel which is not the background, the average value will be closed the real background. However, an effective method to detect that pixel which is not the real background is difficult because the moving pixel point do not know. But through analyzed some fisheye image frames that we know the background is more times than foreground, and the color of foreground is not the same like background. For one pixel in fisheye video sequences, if it is a background pixel, it will be collected around the center of background point in the RGB space for time sample, and the foreground pixel is faraway the center. So we can consider the pixel is not the background which the distance of one color vector to this standard deviation. The calculate formula of standard deviation is

$$S.c = \sqrt{\frac{1}{n-1} \sum_{i=1}^n |X_i.c - X_0.c|^2} \quad (2)$$

Here  $C$  is the three color components of R, G, B,  $X_0$  is the average value of sample. So we can improved above method of background extraction which calculated the standard deviation after average the pixel value of all sample frames, then removed the pixel of average value more than standard deviation. The steps of improved algorithm is

- (1)  $N$  fisheye image frames  $F_i$  was got through sample, here  $i=1,2,\dots,N$ .
- (2) For every pixel  $(x, y)$ , do

a) Calculated the center point

$$F_0(x, y) = \frac{\sum_{i=1}^N F_i(x, y)}{N} \quad (3)$$

b) Calculated the standard deviation

$$S.c(x, y) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N |F_i.c(x, y) - F_0.c(x, y)|^2} \quad (4)$$

c) Calculated the average value of all elements in collection of

$$\{F_i(x, y) \mid \forall c |F_i.c(x, y) - F_0.c(x, y)| \leq S.c(x, y); c = r, g, b; i = 1, 2, \dots, N\} \quad (5)$$

This average value is the background  $B(x, y)$  which closed the real background.

## 2.2 Update the Background

The background will be changed along with the time and the environment, so it needs update automatically. We can divide two kinds of situation for discussions the background changed as the background changes gradually and abruptly.

Case 1, the background changes gradually. Defined a weight function  $w(t)$  as:

$$w(t) = \sum_{x \in W_W} |B(x, t)| \quad (6)$$

For the current time  $t$ , the background is modified as:

$$B_{new}(x) = \frac{1}{2} \left[ \frac{\sum_{t_i=t}^{t-F} G(w(t_i)) F(x, t)}{\sum_{t_i=t}^{t-F} G(w(t_i))} + B_{old}(x) \right] \quad (7)$$

Where  $G(\cdot)$  is the Gaussian function with  $G(0)=1.0$  and  $G(\pm T_1 \times WW)=0.01$ . In real implementation,  $G(\cdot)$  is calculated off-line to generate a look-up table, and the summation in Eq. (7) can be computed iteratively from time  $t-1$  to time  $t$ .

Case 2, the background changes abruptly. We supposed the illumination of the background change satisfies the following linear relation:

$$B_{new}(x) = \alpha B_{old}(x) + \beta \quad (8)$$

And then differentiating both sides with  $x$

$$B'_{new}(x) = \alpha B'_{old}(x) \quad (9)$$

Estimated  $\alpha$  as:

$$\alpha = \sum_{x \in W_W} B'_{new}(x) / \sum_{x \in W_W} B'_{old}(x) \quad (10)$$

During the recovery period, the average difference between the spatial gradient  $F'(x, t)$  of the current image  $F(x, t)$  and that of the old background image  $b(x)$  is calculated as

$$d_g(t) = \frac{1}{W_W} \sum_{x \in W_W} |F'(x, t) - \alpha B'_{old}(x)| \quad (11)$$

Here

$$\alpha_t = \sum_{x \in W_W} F'(x, t) / \sum_{x \in W_W} B'_{old}(x) \quad (12)$$

If there is no object,  $d_g(t)$  should be very small under the linear illumination change model of Eq. (7). Therefore, the image discrimination criterion is defined as:

$$\sigma(t) = \begin{cases} 0, & d_g(t) > Max(D_g) \\ 1, & d_g(t) \leq Max(D_g) \end{cases} \quad (13)$$

Where  $Max(D_g)$  is the predefined maximum gradient difference for image, and the background is renewed as:

$$B_{new}(x) = \frac{\sum_{x \in T_t} F(x, t) \sigma(t)}{\sum_{x \in T_t} \sigma(t)} \quad (14)$$

When the background which closer the true and real time background is extraction, we can calculate the moving blob in fisheye video frames.

### 2.3 Extraction the Moving Blobs Used the Euclidean Distance Method

The moving blobs in the fisheye video sequences can be calculated which used the Euclidean distance algorithm of every frames substrate with background, and then binaryzation used the threshold. In order to get more accurate results, we used the color fisheye video frames to calculate directly, the phasor difference formula of two color pixel used Euclidean distance is

$$\begin{aligned} d &= \sqrt{(x - \mu)^T(x - \mu)} \\ &= \sqrt{(x.r - \mu.r)^2 + (x.g - \mu.g)^2 + (x.b - \mu.b)^2} \end{aligned} \quad (15)$$

Where  $x.r$ ,  $x.g$  and  $x.b$  are the pixel vectors of red, green and blue which corresponded to the three color components respectively, and  $\mu.r$ ,  $\mu.g$ ,  $\mu.b$  are the pixel vectors of background.

We can decide the pixel is the foreground or background in fisheye video frames through the Euclidean distance value  $d$  which compare with the threshold  $T_d$ . If  $d > T_a$ , the pixel is the foreground, else is the background. So selected a suitable  $T_a$  is the key point to detect the foreground moving blobs pixel.

The algorithm steps of extraction the moving blobs is

- (1) Set the background image of  $B(x,y)$ .
- (2) Scan the every pixel of  $(x,y)$  in processing frame, if

$$\begin{aligned} \|F(x, y) - B(x, y)\| \\ = \sqrt{(F(x, y).r - B(x, y).r)^2 + (F(x, y).g - B(x, y).g)^2 + (F(x, y).b - B(x, y).b)^2} \\ \geq T_a \end{aligned} \quad (16)$$

Then output  $G(x,y)=1$ , else  $G(x,y)=0$ .

- (3)  $G(x,y)$  is the moving blob of current frame.

### 2.4 Remove the Shadow of Moving Blobs

The moving object in the fisheye video sequences has some shadow especial sunny day. The shadow is now the part of the moving objects, but the moving blobs of extraction used above algorithm include the shadow because the shadow is moving the same with the moving object. This will make the foreground objects link up into a single stretch, and affect the accuracy of target detection and tracking seriously. In this paper, a shadow remove method was proposed which based on the RGB space.

The shadow and the background are similar in color, but the shadow is darker than the background in brightness. The foreground and the background are different in the chrominance and luminance generally. We have known the color difference of two pixels is shown as the angle difference of pixel vector in RGB space, when the angle

difference is big, the color difference of two pixels is bigger. Set two pixel vector as  $a$ ,  $b$ , then the angle of two pixels is

$$\theta = \arccos \frac{a \bullet b}{\|a\| \times \|b\|} \quad (17)$$

We have known the color is similar of shadow and background, so the angle difference of pixel vector is not big. Define a threshold value, when  $\theta < T_0$  then the color is similar of two pixels. The brightness is the length of pixel vector. Define another threshold value, when the brightness difference of foreground pixel and background pixel is more than  $T_c$ , then this pixel is not the shadow.

But when the background is bright like the foreground, the fixed threshold  $T_0$  will not detect the real shadow. So we need improved this algorithm that get a judge condition in the RGB space which the detect point not changes with the background brightness.

For the vector of  $-(F(x,y)-B(x,y))$ , here  $F(x,y)$  is the pixel vector at  $(x,y)$  in current frame,  $B(x,y)$  is the pixel vector of background at  $(x,y)$ . When the color is similar of two pixel, the angle of  $F(x,y)$  and  $B(x,y)$  is closed. Then the angle of  $-(F(x,y)-B(x,y))$  and  $B(x,y)$  is closer too. So we can calculate the close degree of  $-(F(x,y)-B(x,y))$  and  $B(x,y)$  as

$$\beta = \arccos \frac{-(F(x,y) - B(x,y)) \bullet B(x,y)}{\|F(x,y) - B(x,y)\| \times \|B(x,y)\|} \quad (18)$$

Define the threshold  $T_\beta$ , when  $\beta < T_\beta$ , the color of two pixel is similar. In order not to let the vector  $F(x,y)$  is too small, add another condition of  $\|F(x,y) - B(x,y)\| < T_d$ , here  $T_d$  is another threshold.

The algorithm steps of improved shadow remove in RGB space is

- (1) Get the background image of  $B(x,y)$  and the current frame  $F(x,y)$ .
- (2) For the foreground moving blob pixels  $F(x,y)$  of extracted used the moving blobs extraction algorithm, when

$$\arccos \frac{-(F(x,y) - B(x,y)) \bullet B(x,y)}{\|F(x,y) - B(x,y)\| \times \|B(x,y)\|} < T_\beta \quad (19)$$

$$\|F(x,y) - B(x,y)\| < T_d \quad (20)$$

Then this pixel is the shadow and need remove from the foreground.

## 2.5 Moving Objects Tracking Based on the Moving Blobs

When the moving blobs were detected and the shadows were removed used above algorithms, we can track the moving objects based on the moving blobs used the prediction tracking model. The steps of the moving objects tracking algorithm was designed as

- (1) Extraction the position of moving blobs.

(2) Calculated the area of moving blobs, if the area less than  $T_a$  pixel, then abandon this moving blob. Here  $T_a$  is a threshold of moving blob judgment.

(3) For every moving blob, find the intersection part of predicted position of rectangular and moving blob rectangle in the moving object. If the area of overlap more than half of the object, this blob is belong to the moving object.

1) If only one moving object in accordance with this moving blob, this moving blob is a single moving object.

2) If one moving blob in accordance with more than one moving objects, the moving blob will split and add in every object which the area part is less than the prediction rectangle.

(4) For every moving object which reconstitute based on moving blobs, extracted the minimum frame of this object which include all part of blobs.

(5) Calculated the speed of this moving object.

(6) Predicted the new position of this object.

(7) Checked the moving blob in the frame, if the area of blob was disappeared or less than the threshold  $T_a$  pixel, and then deleted this object in the moving objects list.

(8) If the moving object frame is more than 1/8 of this fisheye image, and the moving blob pixels is less than half of this frame, deleted this object in the moving objects list.

(9) If appear new moving blobs in image frame which can jump the step (2), and judge this moving blob is a new moving object or a part of another moving object. if it is a new moving object, then add a node in the moving objects list table.

(10) Repeated these steps until the program is terminated.

### 3 Experiment Result and Analysis

A fisheye video was captured which the frame rate is 25, and the resolution is 640\*480. We used this video to test all algorithms which based on the moving blob method.

#### 3.1 Background Extraction

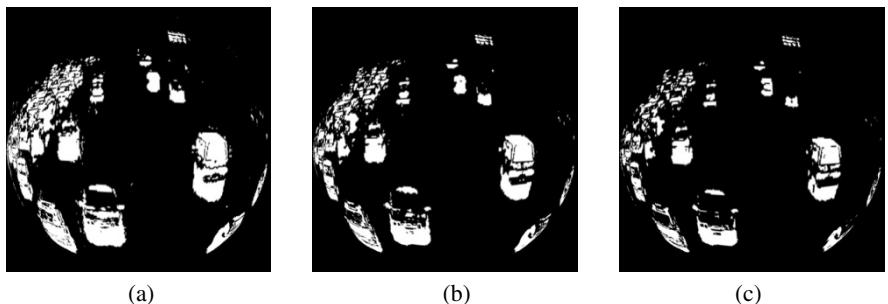
We used the improved mean value algorithm to extract the background. Figure 1(a) is the extraction result which used sample rate 2.5 frames per second, and Figure 1(b) is the extraction result of 5 frames per second. We can know from Figure 1 that the background has a lot of moving mark at the left of image when the video sequences have a lot of moving object because of the low sample rate. The effect became better when increased the sample rate of 5 frames per second in Figure 2. So the sample of the frame rate will affect the background, consider all angles of the question, we think 5 frames per second is enough to extract the background in fisheye video sequences.



**Fig. 1.** The extraction background from fisheye video sequences, (a) is the sample rate of 2.5 frames per second and (b) is the 5 frames per second

### 3.2 The Moving Blob Detection

We used the Euclidean distance to detect the moving blobs in the fisheye video sequences. The experiment results which used difference threshold were shown in Figure 2. The Figure 2(a) was the result of the threshold as 25, it had a lot of adhesions in some moving blobs which it is difficult to separate every moving object, especial at the edge of the fisheye image. The Figure 2(c) was the result of the threshold as 45, the moving blob had not full extracted that means some moving object will discard. The Figure 2(b) was the result of the threshold as 35, we could know that the moving blobs could be detected completely which used this threshold. So we can select this threshold to detect the moving blobs in the fisheye images.

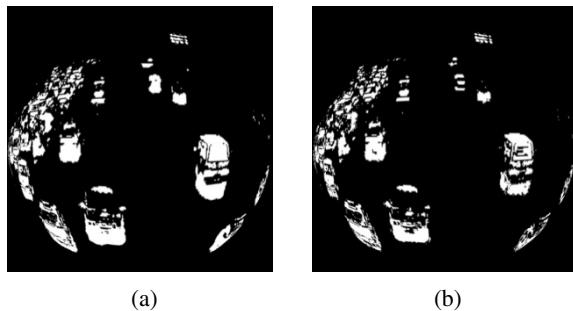


**Fig. 2.** The moving blobs extraction results of difference threshold, (a) is the threshold as 25, (b) is the threshold as 35, and (c) is the threshold as 45

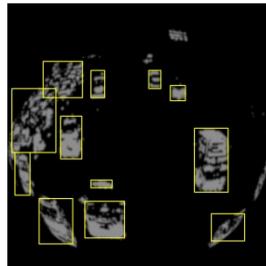
### 3.3 Remove the Shadow of Moving Blobs

From figure 4 we can know it has some interference of shadow for moving objects. It will make some objects connect together. We used the method of improved RGB space shadow remove algorithm, and selected an adapt threshold value of  $T_\beta = 0.1$ ,  $T_d = 60$ , the experiment results after removed the shadow of the moving blobs as shown in Figure 3. The figure 3(a) is include the shadow, and the figure 3(b) is removed the shadow of the moving blobs. This algorithm has a good effect to remove the shadow.

When removed the shadow of the extraction moving blobs, we can mark the moving blobs with rectangular as shown figure 4. Because reduced the adhesion of moving blobs, every blobs can be detected separately.



**Fig. 3.** Removed the shadow of the moving blobs, (a) is include the shadow, and (b) is remove the shadow with the improved algorithm



**Fig. 4.** Mark the blobs of detection



**Fig. 5.** The tracking result of moving object, (a) is one of the original fisheye picture in video sequences, and (b) is the tracking result

### 3.4 Tracking of the Moving Object

The pixel of every blob which be detected was calculated. When the pixels of one blob was less than 80, then discarded this blob as noise. All blobs of the pixels more than 80 were marked at the origin fisheye video sequences. The tracking results of moving

objects were shown in figure 5. From the figure 5 we can know that the moving objects can track stable when it in the center of the picture. When the object is on the edge of the fisheye image, it is difficult to separate because of the big distortion. So this method needs to improve to detect and track the edge object in fisheye image.

## 4 Conclusion

In this paper, we discussed an improved algorithm that used the moving blobs method in the fisheye video sequences. This algorithm divided into two main steps, the first step is detected the moving blobs through the background subtraction, the second step is tracking the moving objects through determine the moving blob. Compared with the traditional background subtraction method, we improved the algorithm of background extraction, the moving blob calculation, and the shadow removed of the moving blobs. The experiment results indicate that the improved algorithm with moving blob method can detect and track some of the moving objects in fisheye video sequences, and it has a good effective when it is in the center of the image. This algorithm also has some shortcomings when too many object is in the edge of the fisheye image, the tracking effect will be weaken because of the big distortion of fisheye image. In our following research work, we will try to establish the distortion model of fisheye image and improve the precision and accuracy of detection and tracking the edge moving object in fisheye video sequences.

**Acknowledgement.** We should like to acknowledge that this work was supported in part by the National Natural Science Foundation (NSFC: 61201435) and the Scientific Research Fund of Hunan Provincial Education Department (Grant No. 13B037). The experiment of this work also was supported in part by the Projects of Hunan Province Science & Technology Department (2013GK3097) and the Research Fund of Key Laboratory of Hunan Province(14k042). Its contents are solely the responsibility of the authors and do not necessarily represent the official views. At the same time, we are thanks to provide lots of experimental support by the Key Laboratory of Optimization and Control for Complex Systems, College of Hunan Province.

## References

- [1] Ying, X., Hu, Z.Y.: Fisheye Lense Distortion Correction Using Spherical Perspective Projection Constraint. Chinese Journal of Computers 26(12), 1702–1708 (2003)
- [2] Huan, Y., Su, H.: A Simple Transforming Model from Fish-eye Image to Perspective Projection Image. Journal of System Simulation 17(1), 29–32 (2005)
- [3] Kannala, J., Brandt, S.S.: A Generic Camera Model and Calibration Method for Conventional, Wide-Angle, and Fish-Eye Lense. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), 1335–1310 (2006)
- [4] Hughes, C., Denny, P., Glavin, M., Jones, E.: Equidistant Fish-Eye Calibration and Rectification by Vanishing Point Extraction. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(12), 2289–2296 (2010)

- [5] Gennery, D.B.: Generalized Camera Calibration Including Fish-Eye Lenses. *International Journal of Computer Vision* 68(3), 239–266 (2006)
- [6] Mitiche, A., Mansouri, A.-R.: On convergence of the Horn and Schunck optical-flow estimation method. *IEEE Transactions on Image Processing* 13(6), 848–852 (2004)
- [7] Wedel, A., Pock, T., Zach, C., Bischof, H., Cremers, D.: An improved algorithm for TV-L1 optical flow. In: *Statistical and Geometrical Approaches to Visual Motion Analysis: International Dagstuhl Seminar, Dagstuhl Castle, Germany*, pp. 23–45 (July 2008)
- [8] Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using Mean Shift. *Computer Vision and Pattern Recognition* (2), 142–149 (2000)
- [9] Qian, Y., Xie, Q.: Camshift and Kalman Predicting Based on Moving Target Tracking. *Computer Engineering and Science* 21(8), 81–83 (2010)
- [10] Yang, B., Zhou, H., Wang, X.: Target Tracking using Predicted CamShift. In: *Proceedings of the 7th World Congress on Intelligent Control and Automation, China*, June 25-27 (2008)

# A Non-negative Low Rank and Sparse Model for Action Recognition

Biyun Sheng<sup>1</sup>, Wankou Yang<sup>1</sup>, Baochang Zhang<sup>2</sup>, and Changyin Sun<sup>1</sup>

<sup>1</sup> School of Automation, Southeast University, Nanjing 210096, China

<sup>2</sup> School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

hisby@126.com, wankou.yang@yahoo.com, bczhang@buaa.edu.cn,  
cysun@seu.edu.cn

**Abstract.** In this paper, we present a new method for video action recognition. The main contributions are two-fold. First, we propose local coordinates contained descriptors (LCCD) instead of appearance-only descriptors. We encode global geometric correspondence by combining descriptors with spatio-temporal locations, which is different from previous methods such as spatio-temporal pyramid matching (STPM). Spatio-temporal location is taken as part of the coding step by utilizing LCCD. Second, a novel non-negative low rank and sparse coding model is developed to encode descriptors for action recognition. Motivated by low rank matrix recovery and completion, local descriptors in a spatio-temporal neighborhood are similar and should be approximately low rank. The objective function is obtained by seeking non-negative low rank and sparse coefficients for local descriptors. The learned coefficients can capture location information and the structure of descriptors, hence improve the discriminability of representations. Experiments validate that our method achieves the state-of-the-art results on two benchmark datasets.

**Keywords:** local coordinates, non-negative low rank, sparse coding, action recognition.

## 1 Introduction

In recent years, action recognition in videos has been a very active research area due to its wide applications such as in surveillance, human-computer interface, sports video analysis, and content based video retrieval [1]. State-of-the-art performances have been achieved by the Bag of Visual Words (BOVW) method, which includes extraction of local descriptors (e.g., HOG or HOF) and construction of representations.

In the framework of BOVW, the collection of unordered words ignores the interest points' location information. Aiming at the loss of location information, Choi et al. extend the spatial pyramid method for video retrieval and propose Spatio-Temporal Pyramid Matching (STPM) [6]. The concatenation of histograms leads to huge vector representation. The finer the region is portioned, the longer the final representation is. Yuan et al. introduce a new global

feature called 3D R transform, which captures the distribution of interest points [1]. The global feature and the BOVW representation are then combined by a context-aware feature fusion method. The method improves the accuracy while it brings computational complexity.

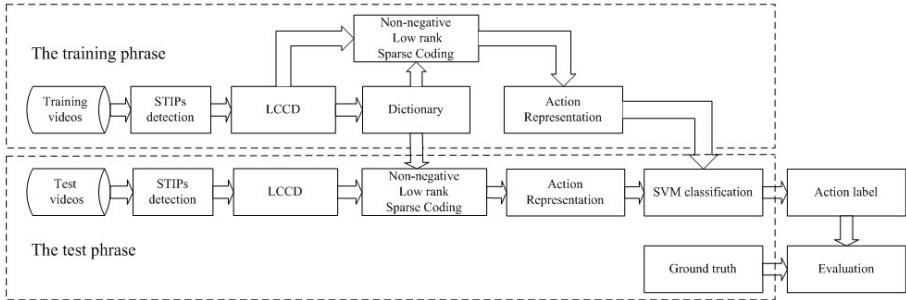
Restrictive cardinality constraint on vector quantization in BOVW leads to relatively high reconstruction error. Sparse coding technique has attracted much attention to reduce the reconstruction error. However, it has a drawback that sparse codes cannot vary smoothly on the data manifold. The dependence of local descriptors is ignored which results in different codes for similar descriptors. Gao et al. propose Laplacian sparse coding to exploit the dependence among the local descriptors [2]. This algorithm preserves the consistence of sparse representation for similar local descriptors while the large number of descriptors leads to computational infeasibility as well as impracticality in real-world applications.

In fact, local spatially and temporally descriptors close in a video should have similar sparse codes ideally. The low rank representation can easily solve the non-consistency problem [4]. Promising results have been shown by low rank and sparse matrix recovery in many applications [3],[8]. However, limited work has applied the low rank sparse coding framework to solve action recognition problem.

Usually sparse coding technique or its variant is followed by max pooling to get the final representation. The sign of coding coefficients is not constrained traditionally. Negative coefficients appear in order to satisfy the objective function, while large numbers of zero coefficients are inevitable. Since non-zero components typically provide useful information, the max pooling process will bring the loss in terms of those negative components, and further degrade the classification performance [5]. Besides, it is meaningful to reduce the information loss by non-negative constraint on coding coefficients during the encoding process.

In this paper, we propose new descriptors called local coordinates contained descriptors (LCCD) and calculate corresponding coefficients by non-negative low rank sparse coding method. Fig.1 shows the flowchart of our framework. We encode coordinates of spatio-temporal interest points (STIPs) as well as the corresponding descriptors so that the representations themselves contain location information. For the encoding model, we add low rank regularizer and non-negative constraint into the traditional sparse coding objective function. The low rankness enforces similar descriptors to have similar sparse codes, which considers the local geometrical structure of the data. The non-negative constraint lowers information loss for representations. Therefore, the learned representations are remarkably more discriminative.

The remainder of this paper is organized as follows. Section 2 introduces the new descriptors with appearance feature and location information, called LCCD. Section 3 presents the non-negative low rank and sparse coding method. Section 4 experimentally tests our method on two human action datasets. Section 5 concludes the paper.



**Fig. 1.** Flowchart of the proposed action recognition framework

## 2 Local Coordinates Contained Descriptors (LCCD)

For a video, a set of interest points are detected and traditional descriptors are obtained based on every interest point. These descriptors own the local texture or motion information, however, ignores the location information of the interest points. Instead of utilizing more complicated descriptors such as 3D R transform [1] or settling the problem during pool stage by STPM [6], we propose a more intuitive and easier method to describe video descriptors, namely local coordinates contained descriptors (LCCD).

We first perform STIPs detection by the Harris operator. A multi-scale approach is used. The HOG/HOF feature is adopted to describe the cuboid extracted at each interest point [12]. LCCD of a video are denoted as:

$$\begin{cases} X = [X_1, X_2, \dots, X_i, \dots, X_N] \\ X_i = [\varphi_i; \alpha x_i; \alpha y_i; \beta t_i] \end{cases}, \quad 1 \leq i \leq N \quad (1)$$

where  $\alpha$  and  $\beta$  are parameters with functions of coordinate normalization, location weight regulation and dimensional transformation,  $(x_i, y_i, t_i)$  is the coordinate of the  $i^{th}$  interest point,  $\varphi_i$  is the HOG/HOF feature, and  $N$  is the total number of interest points detected in the video.

In contrast to the original appearance-only descriptors, the proposed ones contain location information which is beneficial to capture geometric structure of the data. Compared with STPM, there is no need for dividing the video artificially to define the pooling regions. Appearance-only descriptors and their coordinates are simultaneously encoded so that the learned coefficients have more discriminative power.

## 3 Non-negative Low Rank Sparse Coding

In this section, we first introduce the non-negative low rank sparse model and then give the optimization process.

### 3.1 Non-negative Low Rank Sparse Model

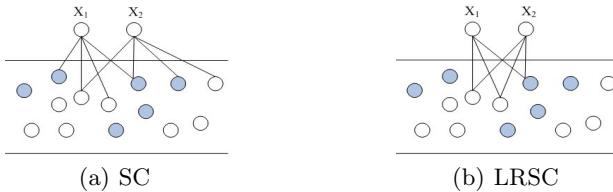
In the sparse model, the input signal is well approximated by a sparse linear combination of the given overcomplete bases in dictionary. Such sparse representations are usually derived by linear programming as an  $l_1$ -norm minimization problem. But the  $l_1$  based regularization is sensitive to outliers. Therefore, we use  $l_{2,1}$ -norm instead in this paper.

Suppose  $X = [X_1, X_2, \dots, \dots, X_N] \in \mathbb{R}^{d \times N}$  be LCCD for a video, in which  $d$ ,  $N$  respectively denote the dimension and number of descriptors. The proposed non-negative low rank sparse model is

$$\min_U \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|U\|_* + \lambda_2 \|U\|_{2,1} \quad s.t. \ U \geq 0 \quad (2)$$

where  $\|\cdot\|_F$ ,  $\|\cdot\|_*$ ,  $\|\cdot\|_{2,1}$  respectively denotes the Frobenius-norm, the nuclear norm, and the  $l_{2,1}$ -norm of a matrix.  $U = [U_1, U_2, \dots, U_N]$  is the coefficient matrix with each  $U_i$  being the representation of  $X_i$ . The nuclear norm, a convex approximation to the rank function, is the sum of the singular values of a matrix.  $\lambda_i (i=1, 2)$  are parameters to trade off low rankness and sparsity.

From the proposed model in (2), we can find the fact that the model degenerates to the sparse coding model if we set the parameter  $\lambda_1 = 0$ . The nuclear norm here is used to enforce the codes of similar descriptors in neighborhood to be approximately similar. Fig.2 shows the comparison between standard sparse coding (SC) and our low rank sparse coding (LRSC). Different from SC, similar bases are selected to guarantee the consistency of similar descriptors in LRSC.



**Fig. 2.** Comparison between SC and LRSC.  $X_1$  and  $X_2$  are two similar inputs to be encoded.

Without non-negative constraint, the coefficients learned by low rank sparse model can be negative. Zero (or small positive) coefficients indicate the corresponding bases in the dictionary have no (or very small) influence. However, since zero (or positive value) is always larger than negative values, max pooling strategy will choose zero (or positive value) instead of negative values [5]. It not only leads to worse performance for data representation, but also lacks physical interpretation for many visual data. Therefore, the non-negative constraint on the coefficients is meaningful and necessary.

### 3.2 Optimization Process

Inexact Augmented Lagrange multipliers (IALM) have been applied to solve the low rank problem [8]. We first introduce two auxiliary variable V and W to make regularizations of the objective function in (2) separable. The problem (2) can be transformed as follows:

$$\min_{U,V,W} \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} \quad s.t. \quad U = V, U = W, W \geq 0 \quad (3)$$

The augmented Lagrangian function of problem (3) is

$$\begin{aligned} L(U, V, W, Y_1, Y_2, u) &= \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} + \langle Y_1, U - V \rangle \\ &\quad + \langle Y_2, U - W \rangle + \frac{u}{2} (\|U - V\|_F^2 + \|U - W\|_F^2) \\ &= \frac{1}{2} \|X - BU\|_F^2 + \lambda_1 \|V\|_* + \lambda_2 \|W\|_{2,1} \\ &\quad + h(U, V, W, Y_1, Y_2, u) - \frac{1}{2u} (\|Y_1\|_F^2 + \|Y_2\|_F^2) \end{aligned} \quad (4)$$

where

$$\begin{cases} h(U, V, W, Y_1, Y_2, u) = \frac{u}{2} (\|U - V + \frac{1}{u} Y_1\|^2 + \|U - W + \frac{1}{u} Y_2\|^2); \\ \langle A, B \rangle = \text{tr}(A^T B) \end{cases} \quad (5)$$

The dictionary B in (3) is calculated by k-means. By the method of IALM, the objective function achieves convergence by a sequence of closed form update steps. The variable U, V or W is updated with other variables fixed. The updating schemes are as follows.

$$\begin{aligned} U_{k+1} &= \underset{U}{\operatorname{argmin}} \frac{1}{2} \|X - BU\|_F^2 + \langle Y_{1,k}, U - V_k \rangle \\ &\quad + \langle Y_{2,k}, U - W_k \rangle + \frac{u_k}{2} (\|U - V_k\|_F^2 + \|U - W_k\|_F^2) \\ &= (B^T B + 2u_k I)^{-1} (B^T X - Y_{1,k} - Y_{2,k} + u_k V_k + u_k W_k) \end{aligned} \quad (6)$$

$$\begin{aligned} V_{k+1} &= \underset{V}{\operatorname{argmin}} \frac{\lambda_1}{u_k} \|V\|_* + \frac{1}{2} \|V - (U_k + \frac{1}{u_k} Y_{1,k})\|_F^2 \\ &= \Theta_{\frac{\lambda_1}{u_k}} (U_k + \frac{1}{u_k} Y_{1,k}) \end{aligned} \quad (7)$$

$$\begin{aligned} W_{k+1} &= \underset{W \geq 0}{\operatorname{argmin}} \frac{\lambda_2}{u_k} \|W\|_{2,1} + \frac{1}{2} \|(W - (U_k + \frac{1}{u_k} Y_{2,k}))\|_F^2 \\ &= \max(\Omega_{\frac{\lambda_2}{u_k}} (U_k + \frac{1}{u_k} Y_{2,k}), 0) \end{aligned} \quad (8)$$

where  $\Theta$  and  $\Omega$  are respectively singular value soft-thresholding operator and  $l_{2,1}$  minimization operator. In detail, the form of analytic solutions for  $\Theta$  and  $\Omega$  are as follows:

$$\Theta_\lambda(A) = U_A S_\lambda(\Sigma_A) V_A^T \quad (9)$$

In (9),  $A = U_A \Sigma_A V_A^T$  is the SVD of A and  $S_\lambda(A_{ij}) = sign(A_{ij})max(0, |A_{ij}| - \lambda)$  is soft-thresholding operator.

Let  $A = [a_1, a_2, \dots, a_i, \dots]$  be a given matrix, then the  $i^{th}$  column of  $\Omega_\lambda(A)$  is  $\frac{\max(0, \|a_i\| - \lambda)}{\|a_i\|} a_i$ .

*Algorithm1. Non-Negative Low Rank Sparse Coding via IALM*

```

Input:Data X, Dictionary B, and Parameters  $\lambda_1$  and  $\lambda_2$ ;
Output:U,V,W;
const
     $\rho = 1.1; u = 0.1; maxiter = 10e30; \varepsilon = 10e-3;$ 
var
    iter: 0..maxiter;
begin
    iter := 0;
repeat
    fix V,W and update variable U according to (6);
    fix W,U and update variable V according to (7);
    fix U,V and update variable W according to (8);
     $Y_{1,iter+1} := Y_{1,iter} + u(U_{iter} - V_{iter});$ 
     $Y_{2,iter+1} := Y_{2,iter} + u(U_{iter} - W_{iter});$ 
     $u = \rho u; \quad iter := iter + 1;$ 
until  $\|U - V\|_\infty < \varepsilon$  and  $\|U - W\|_\infty < \varepsilon$ ;or iter = maxiter;
end.
```

## 4 Experiments

We test our approach on two benchmark datasets: the KTH actions dataset [13], and the UCF Sports dataset [14].

### 4.1 Experiments on the KTH Dataset

The KTH action dataset contains six types of human actions (boxing, hand waving, hand clapping, walking, jogging and running), performed repeatedly by 25 subjects in four different scenarios: outdoors, outdoors with camera zoom, outdoors with different clothes and indoors. Twenty-four actors' videos are used as the training sets and the remaining one person's videos as the testing set. The results are the average of 25 times runs. We empirically set the size of the dictionary to 250 for the dataset. For the non-negative low rank sparse model, we set the tradeoff parameters  $\lambda_1 = 1, \lambda_2 = 0.1$ .

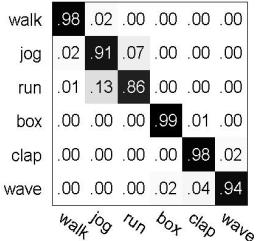
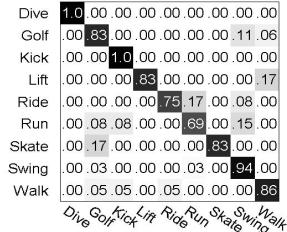
**Fig. 3.** Confusion matrix on KTH**Fig. 4.** Confusion matrix on UCF

Fig.3 shows the confusion matrix across all scenarios. The figure demonstrates the effectiveness of our proposed approach. For example, the accuracies of some actions such as "walking", "boxing" and "handclapping" can reach above 97%. The "running" action is easily misclassified as "jogging" because of the high similarity between the two actions. Table 1 lists the average accuracy of action recognition by other researchers.

**Table 1.** Comparison with previous work on the KTH dataset

Approach	Year	Accuracy(%)
Brendel et al.[17]	2010	94.22
Le et al.[7]	2011	93.90
Zhang et al.[16]	2012	95.5
Wang et al.[15]	2013	94.2
<b>Ours</b>		<b>94.32</b>

Compared with the listed results in recent research, our method achieves 94.32%, which is comparable to the state-of-the-art result. However, with the original appearance-only descriptors the recognition rate is 93.32%. The experimental result illustrates that the proposed descriptors improve the performance of our framework. Traditional descriptors don't utilize the location information of STIPs (or settle the problem during pooling stage), which leads to a set of unordered representations (or lengthy representations). Our descriptors contain location of the interest point and appearance characteristics of cuboid around the point.

In order to validate the effectiveness of non-negative constraint, low rank and  $l_{2,1}$ -norm regularizer, we change one of the above three terms with others fixed. The results of comparison are shown in Table 2. When we calculate absolute values of coefficient matrix instead of non-negative constraint, the accuracy is only 91.82%. It shows that taking absolute values artificially is not reasonable in our model and may drop the accuracy significantly. The performance of our

**Table 2.** Effectiveness of non-negative low rank sparse model

Method	Accuracy(%)
Without non-negative constraint	91.82
Without low rank regularizer	93.32
$l_1$ -norm instead of $l_{2,1}$ -norm	93.82
<b>Ours</b>	<b>94.32</b>

model without low rank regularizer is 93.32%. If we change  $l_{2,1}$ -norm to  $l_1$ -norm, the accuracy is 93.82% . In combination of the three terms, the accuracy of our method is 94.32%. The experimental results demonstrate that the representations obtained by our proposed method are more discriminative.

## 4.2 Experiments on the UCF Sports Dataset

The UCF Sports dataset consists of 150 videos with 9 action classes taken from real broadcasts (e.g., diving, golf swinging, kicking), with different viewpoints and scene backgrounds. The dataset is tested in a leave-one-out manner, cycling each example in as a test video one at a time. We empirically set the size of the dictionary to 800 for the dataset. For the non-negative low rank sparse model, we set the tradeoff parameters  $\lambda_1 = 1, \lambda_2 = 0.1$ .

Fig.4 shows the confusion matrix of our approach on the UCF dataset. The recognition rate for some actions is high up to 100% such as "Dive" and "Kick". Experimental results by previous methods are listed in Table 3.

**Table 3.** Comparison with previous work on the UCF dataset

Approach	Year	Accuracy(%)
Kovashka et al.[18]	2010	87.27
Le et al.[7]	2011	86.5
Yuan et al.[1]	2012	87.33
Wang et al.[15]	2013	88.0
<b>Ours</b>		<b>88.0</b>

When we use the traditional descriptors and pool the coefficients by STPM, the accuracy is about 80% which is much lower than our method. The result shows that the proposed LCCD is especially fit for the UCF dataset. We do the same experiments as in section 4.1 to validate our proposed model, and Table 4 illustrates the performances of different combinations. Non-negative constraint here is vital and effects the final result largely.

**Table 4.** Effectiveness of non-negative low rank sparse model

Method	Accuracy(%)
Without non-negative constraint	82.0
Without low rank regularizer	86.67
$l_1$ -norm instead of $l_{2,1}$ -norm	88.0
<b>Ours</b>	<b>88.0</b>

## 5 Conclusion

In this paper, we have presented a novel method to learn representations of human actions. In order to describe the "where" property of STIPs, we encode descriptors with location information. Besides, we adopt non-negative low rank sparse coding technique. The learned coefficients have the property of spatio-temporal consistency and finally boost the accuracy. Extensive experiments on two datasets have demonstrated the effectiveness of our proposed approach.

**Acknowledgments.** TThis work is supported by National Natural Science Foundation (NNSF) of China under Grant . 61375001, 61473086, partly supported by the open fund of Key Laboratory of Measurement and partly supported by Control of Complex Systems of Engineering, Ministry of Education (No. MCCSE2013B01), and the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), (No. 30920130122006).

## References

- Yuan, C.F., Li, X., Hu, W.M., Ling, H.B., Maybank, S.: 3D R Transform on Spatio-Temporal Interest Points for Action Recognition. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Portland (2013)
- Gao, S.H., Tsang, I.W.H., Chia, L.T., Zhao, P.L.: Local features are not lonely - laplacian sparse coding for imageclassification. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–2. IEEE Press, San Francisco (2010)
- Zhuang, L.S., Gao, H.Y., Lin, Z.C., Ma, Y., Zhang, X.: Non-Negative Low Rank and Sparse Graph for Semi-Supervised Learning. In: 25th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2328–2331. IEEE Press, Providence (2012)
- Zhang, T.Z., Ghanem, B., Liu, S., Xu, C.S., Zhang, X., Yu, N.H., Ahuja, N.: Low-Rank Sparse Coding for Image Classification. In: 14th IEEE International Conference on Computer Vision, pp. 281–286. IEEE Press, Sydney (2013)
- Zhang, C.J., Liu, J., Tian, Q., Xu, C.S.: Image Classification by Non-Negative Sparse Coding, Low-Rank and Sparse Decomposition. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1673–1678. IEEE Press, Colorado (2011)

6. Choi, J., Wang, Z.Y., Lee, S.C.: Spatio-temporal pyramid matching for sports videos. In: Proceeding MIR 2008 Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 291–297. IEEE Press, New York (2008)
7. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3361–3368. IEEE Press, Providence (2011)
8. Zhang, Y.M.Z., Jiang, Z.L., Davis, L.S.: Learning Structured Low-rank Representations for Image Classification. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–3. IEEE Press, Portland (2013)
9. Jiang, Z.L., Ghanem, B., Liu, S., Ahuja, N.: Low-rank sparse learning for robust visual tracking. In: 12th European Conference on Computer Vision, pp. 470–474. IEEE Press, Florence (2012)
10. Zhang, Z.D., Matsushita, Y., Ma, Y.: Camera Calibration with Lens Distortion from Low-rank Textures. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2321–2328. IEEE Press, Providence (2011)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 19th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, New York (2006)
12. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning Realistic Human Actions from Movies. In: 21th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Alaska (2008)
13. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Proceedings of International Conference on Pattern Recognition, 17th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–4. IEEE Press, Washington (2004)
14. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 21st IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7. IEEE Press, Alaska (2008)
15. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Dense Trajectories and Motion Boundary Descriptors for Action Recognition. IJCV 103, 60–79 (2013)
16. Zhang, Y.M., Liu, X.M., Chang, M.C., Ge, W.N., Chen, T.: Spatio-Temporal Phrases for Activity Recognition. In: 12th European Conference on Computer Vision, pp. 707–721. IEEE Press, San Francisco (2012)
17. Brendel, W., Todorovic, S.: Activities as Time Series of Human Postures. In: 11th European Conference on Computer Vision, pp. 9–13. IEEE Press, Greece (2010)
18. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 23rd IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–4. IEEE Press, San Francisco (2010)

# Extreme Learning Machine Based Hand Posture Recognition in Color-Depth Image

Zhen Zhou, Shutao Li, and Bin Sun

College of Electrical and Information Engineering,

Hunan University, Changsha, China, 410082

icyray828@gmail.com, shutao.li@hnu.edu.cn, sunbinxs@126.com

**Abstract.** Hand posture recognition is one of the most challenging problems in the computer vision field, especially in the scenes with complex background and illumination variance. This paper presents a real time hand posture recognition method in color-depth image. To accurately locate hands in the images with complex background, a depth histogram based adaptive thresholding method is adopted for the depth image and a Bayesian skin-color detection is performed for the corresponding color image. Then two processed results are fused and refined with a region-growing method. Finally, the histogram of gradients feature of the hand posture is computed for Extreme Learning Machine classifier to recognize different postures. Experiments show that the proposed hand posture recognition method runs in real-time and achieves high recognition accuracy.

**Keywords:** Hand detection, posture recognition, Extreme Learning Machine, HOG feature extraction.

## 1 Introduction

Vision based hand posture recognition has become a popular research domain in the computer vision field. It plays an important role in applications such as human-computer interaction (HCI)[1], robotic design[2] and so on.

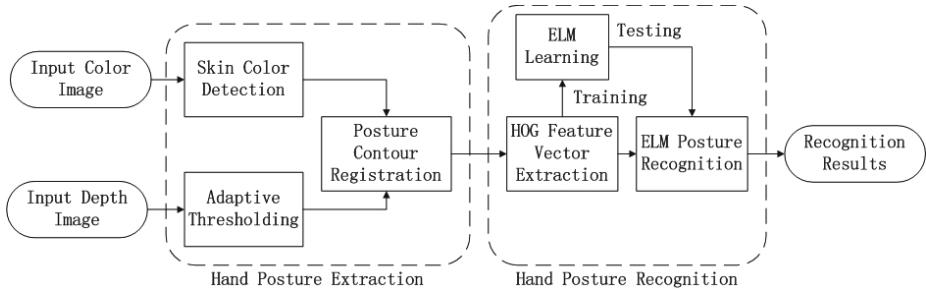
Generally, vision based hand posture recognition methods can be divided into two categories: static hand posture recognition and dynamic hand gesture recognition. For dynamic gesture recognition, a series of the hand motion trajectory is detected for recognition. Elmezain *et al.*[3] proposed a gesture trajectory recognition method based on orientation dynamic feature and Hidden Markov Models (HMM). For static posture recognition, kinds of visual characteristics of single frame is extracted for recognition. Weng *et al.*[4] took advantage of color-cue and velocity weighted feature for hand segment and adopted density distribution feature for posture recognition. Gorce *et al.*[5] proposed a 3D generative hand model and classified hand postures by seeking the minimum residual between the parametric model and the observed image. Since advances in commodity-level RGB-D sensor such as Kinect have greatly simplified the problem of capturing

depth data, some hand posture recognition methods based on color-depth image are also proposed. Bagdanov *et al.*[6] located hands with a predicted depth thresholding algorithm and extracted SURF features for posture recognition.

In this paper, a new hand posture recognition method is proposed for color-depth image. The proposed method is comprised of two blocks: hand posture extraction and hand posture recognition. Depth histogram based adaptive thresholding and Bayesian skin-color detection are combined to extract hands in complex background and illumination conditions, then a region-growing based contour registration method is adopted to refine the extracted hand posture. Finally a robust recognition algorithm using the HOG feature and extreme learning machine (ELM) is proposed to recognize hand postures.

## 2 The Proposed Hand Posture Recognition Method

Fig.1 is the overview of the proposed hand posture recognition method. It consists of two parts, i.e., hand posture extraction and hand posture recognition. For hand posture extraction, a depth histogram based adaptive thresholding method is proposed to get rid of body and background in the depth image, and Bayesian classifier based skin color detection is performed to locate skin-like area in the corresponding color image. Then the two results are combined to locate the hand posture region. To overcome the unmatched edges and artifacts of color-depth image, a region growing based contour registration method is adopted. For hand posture recognition, the HOG feature is extracted and used for training of the ELM classifier. Finally the trained ELM classifier is used to recognize the posture in each input RGB-D image.

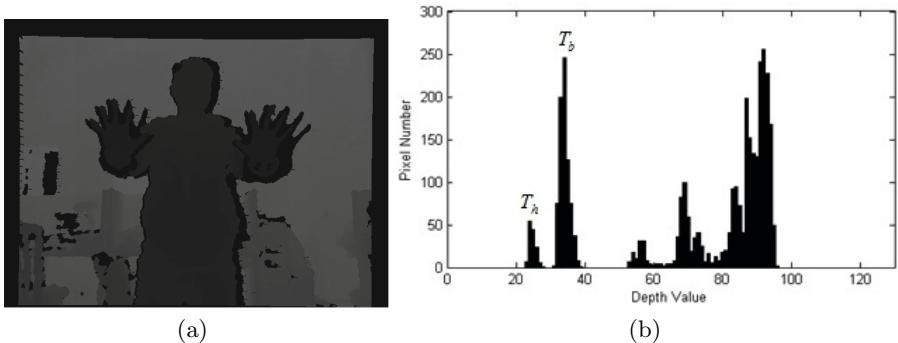


**Fig. 1.** Overview of the proposed hand posture recognition method

### 2.1 Hand Posture Extraction

For hand posture extraction, the skin color region is detected on the input color image and adaptive thresholding is performed on the input depth image. A hand posture image can be then obtained by combining the two results and refined with an region-growing based contour registration algorithm.

An adaptive thresholding for the depth histogram of the input depth image is used to locate hand region. Generally, human body is always in front of camera and hands are always in front of body during the human-computer interaction by RGB-D sensor[7]. So body and background generate two peaks with the largest number of pixels in the depth histogram as shown in Fig.2. The one with smaller depth value is considered to be the body depth value  $T_b$ . Then the depth value of hands  $T_h$  is defined by the peak with lower depth value than  $T_b - d$ , where  $d$  is a distance value. The depth value with the smallest number of pixels between  $T_h$  and  $T_b$  is used as the depth threshold to obtain the hand region in the depth image.



**Fig. 2.** Example of depth image and its depth histogram, (a) the depth image, (b) the corresponding depth histogram

Skin color detection is performed with a Bayesian approach[8] for the input color image. The input color image is transformed into the YUV color. Since the Y component is very sensitive to the illumination variation, only the (U, V) components are used for skin color detection. Denoting the input pixel color as  $c$ , the skin color as  $s$  and the non-skin color as  $n$ , the probability to be a skin color pixel for the input color pixel be obtained as follows

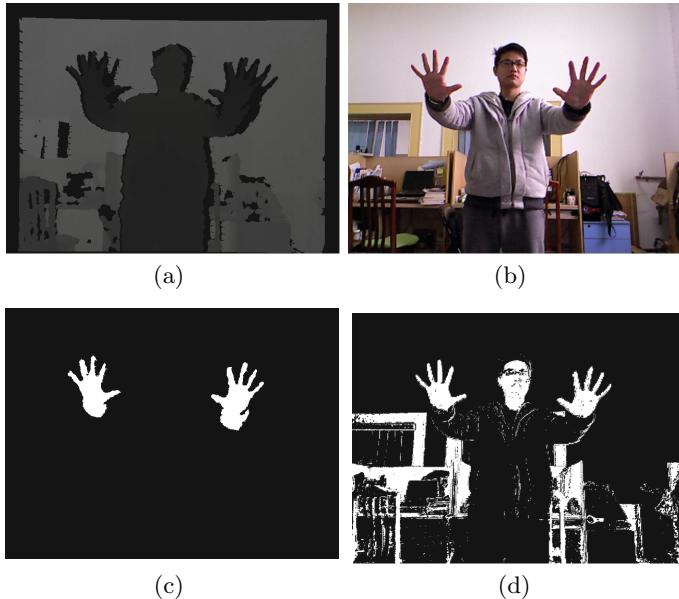
$$P(s|c) = \frac{P(c|s)P(s)}{P(c|s)P(s) + P(c|n)P(n)} \quad (1)$$

Then the input pixel with color  $c$  is considered to be a skin pixel if  $P(s|c)$  is greater than a threshold value  $T$ .

Fig.3 shows the results of adaptive depth thresholding and skin color detection. After skin color detection and depth thresholding, the hand region is obtained by combining the two results together as

$$M_r(x, y) = M_d(x, y) \text{ AND } M_s(x, y) \quad (2)$$

where  $M_r(x, y)$  is the binary hand region mask,  $M_d(x, y)$  and  $M_s(x, y)$  are the results of depth thresholding and skin color detection respectively.



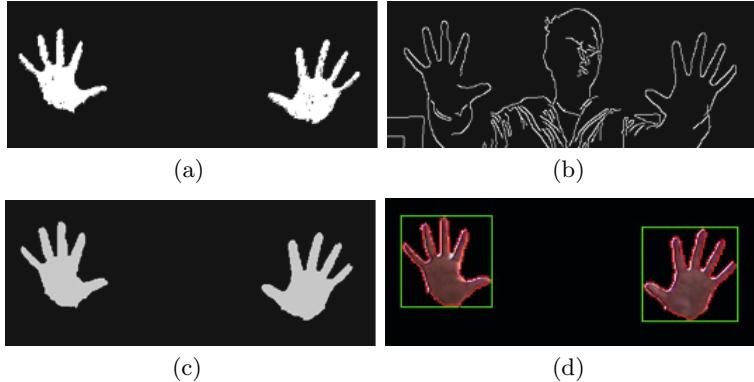
**Fig. 3.** Example of adaptive depth thresholding and skin color detection, (a) original depth image, (b) original color image, (c) result of depth thresholding, (d) result of skin color detection

Due to unmatched edges and invalid pixels in the RGB-D image, the obtained hand region  $M_r$  may be inaccurate. To obtain the accurate hand posture region, a region-growing based method[9] is used to align the contours of the hand region and color image. A region grows from contours of rough posture image until it reaches to the nearest edge or a certain distance. The accurate hand posture region is obtained by

$$M = M_r \text{ } XOR \text{ } M_a \quad (3)$$

where  $M$  is the refined hand posture region,  $M_r$  is the rough posture region and  $M_a$  is the amendment region.

Fig.4 shows an example of posture contour registration results. Compared with the rough posture image in Fig.4(a), the refined posture region in Fig.4(c) has much better contours. Then precise hand postures are obtained by combining the refined posture region and the color image. The hand posture image is extracted by segmenting each posture from the obtained hand posture region in a bounding box, as shown in Fig.4(d).



**Fig. 4.** Example of hand posture contour registration, (a) rough hand posture region, (b) result of edge detection, (c) refined hand posture regions after registration, (d) extracted hand postures

## 2.2 Hand Posture Recognition

For hand posture recognition, the HOG feature of the extracted hand posture image is computed and used for the ELM classifier to recognize different hand postures.

To extract the HOG features of the hand posture, the gradient magnitude and gradient orientation of each pixel are firstly computed as follows

$$\begin{cases} G_x(x, y) = I(x + 1, y) - I(x - 1, y) \\ G_y(x, y) = I(x, y + 1) - I(x, y - 1) \end{cases} \quad (4)$$

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \quad (5)$$

$$\Phi(x, y) = \tan^{-1}(G_y(x, y)/G_x(x, y)) \quad (6)$$

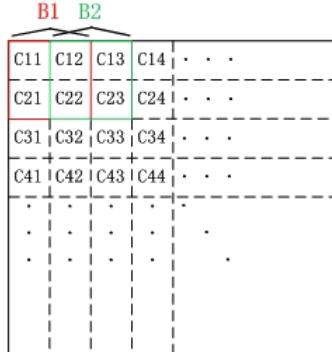
where  $I(x, y)$  is the pixel value at  $(x, y)$ ,  $G_x(x, y)$  and  $G_y(x, y)$  are the horizontal and vertical direction gradients respectively,  $G(x, y)$  denotes the gradient magnitude, and  $\Phi(x, y)$  denotes the gradient orientation.

After the gradient magnitude image and the gradient orientation image are computed, they are divided into  $n_{bl}$  pixel blocks. Then each block is divided into  $n_{ce}$  pixel cells, and the gradient orientation range is also divided into  $n_{bi}$  bins for each cell. The gradient histogram of a cell is computed by adding the magnitude of each pixel in the cell to the corresponding orientation bin with a global weight  $\theta$ . The weight of each pixel in the posture image is calculated by a two-dimensional Gaussian function as follows

$$\theta = \exp(-((x - w/2)^2 + (y - h/2)^2)/(2 * \delta^2)) \quad (7)$$

where  $w$  and  $h$  denote the width and the height of the hand posture image, and  $\delta$  denotes a scale factor. Then the histogram of block  $H_B$  is generated by

concatenated histograms of cell and normalized afterwards. The HOG feature of the whole image is generated by concatenating all the gradient histograms of different blocks together and its dimension is  $n_{bl} \times n_{ce} \times n_{bi}$ . Fig.5 shows the construction of the HOG feature vector.



**Fig. 5.** Construction of HOG feature vector

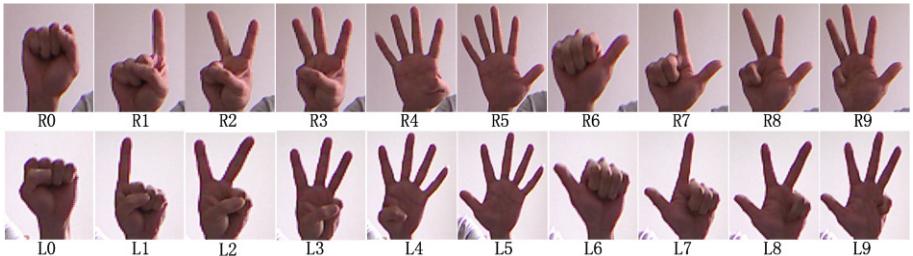
After the hand posture's HOG feature vector is generated, an ELM[11] classifier can be constructed. ELMs was developed for the single-hidden-layer feedforward neural networks (SLFNs) at the first time and then turn into the generalized SLFNs. It is adopted for hand posture recognition because of its fast learning speed, concise human intervene and varietal computational scalability[12].

In the training process, HOG feature vectors of the posture images and their corresponding class labels are sent to ELM as a training set. The hidden node number  $L$  is tuned and sigmoid function is selected as the activation function of ELM. Then a model with optimal network parameters are generated by training. In the testing process, class labels of testing HOG feature vectors are predicted using the trained model.

### 3 Experiments

In order to evaluate the proposed hand posture recognition method, we collected 12,000 color-depth images with the Microsoft Kinect sensor. The images are taken from 10 different persons and contain 20 different postures. Each posture has 600 color-depth images. Fig.6 shows the defined 20 postures to be recognized in the experiments. The proposed hand recognition method is implemented with C++ and the experiments are conducted on a PC with a 2.53 GHz CPU and 2 GB RAM. The parameter values of the proposed method used in the experiments are given in Table 1.

Offline hand posture recognition experiment is performed by randomly choosing 200 pairs of images for each posture for training and using the other 400 pairs

**Fig. 6.** the defined 20 postures**Table 1.** Parameter values of experiments

Step	Parameter	Value
Hand Posture Extraction	Distance value $d$	4
	Skin-color threshold value $T$	0.1
Hand Posture Recognition	Scale factor $\delta$	4.0
	Block number $n_{bl}$	49
	Cell number $n_{ce}$	4
	Bin number $n_{bi}$	9
	Hidden nodes $L$	150

for testing. The test is repeated for 5 times and the average value is considered as the result. The comparing evaluation is also conducted for ELM and SVM to compare their accuracy rate and time cost.

Table 2 shows the results of the proposed posture recognition method on offline dataset. It can be found that ELM based method achieves an accuracy rate of 98.05%, whereas SVM based method reaches 93.88%. The hidden layer of ELMs needs not to be tuned, the input weights and hidden layer biases of ELMs can be randomly assigned and only a small amount of iterative processes are needed. These merits make ELM performs significantly better and faster than the active parameters selected SVM on the classification. Results show that the recognition rate of posture L0 and R0 is a bit lower than other postures, and the reason may be that the posture L0 and R0 have similar contours and similar central parts with the highest weighted when the HOG features are computed.

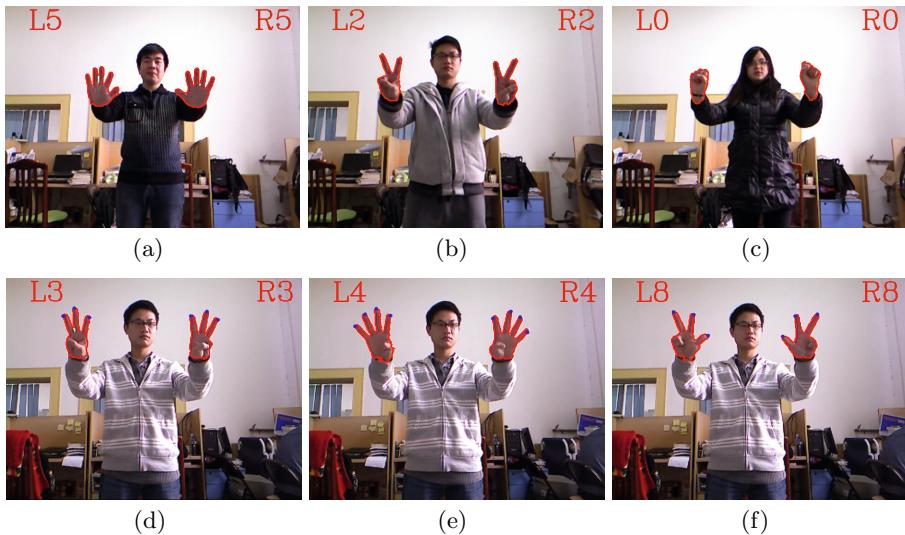
Online testing is also performed with a Kinect sensor. A sequence of 1,870 frames with a resolution of 640\*480 is tested. All the 20 different postures occur in the testing sequence. Experiment results show the proposed hand recognition method runs about 23 frames per second and achieves an accuracy rate of 96.03%. Fig.7 shows the examples of the online posture recognition results.

To verify the robustness of the proposed hand recognition method to illumination variation, RGB-D images with the different illuminations are collected and recognized with the proposed method. Fig.8 shows the examples of the

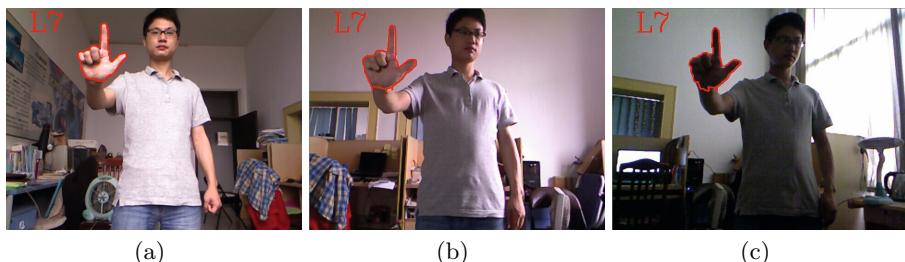
hand posture recognition results. Although the three images in Fig.8 have significant illumination difference, the proposed method is also able to recognize the

**Table 2.** Hand posture recognition rate

Posture	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	R0
ELM(%)	93.50	99.25	97.00	97.25	98.50	99.50	99.00	98.25	98.50	98.00	94.75
SVM(%)	85.75	92.00	92.50	93.50	95.25	98.50	97.25	96.50	94.25	94.00	87.00
Posture	R1	R2	R3	R4	R5	R6	R7	R8	R9	Average	
ELM(%)	98.75	98.75	98.25	98.00	99.00	99.75	99.50	97.50	97.25	98.05	
SVM(%)	94.50	95.25	92.75	93.00	97.25	96.50	97.25	93.50	94.00	93.88	



**Fig. 7.** Examples of the online posture recognition results, (a) L5 and R5, (b) L2 and R2, (c) L0 and R0, (d) L3 and R3, (e) L4 and R4, (f) L8 and R8



**Fig. 8.** Examples of the hand posture recognition results with illumination variation, (a) bright illumination, (b) normal illumination, (c) dark illumination

posture in them correctly. Experiment results show that the proposed hand posture recognition method is robust to the illumination variation.

## 4 Conclusions

In this paper, a new real-time hand posture recognition method is proposed. A depth histogram based adaptive thresholding method is adopted to locate hands in the input depth image, and the skin color region is detected in the corresponding color image by using the Bayesian rule. Then the results are combined and refine by a region growing algorithm to obtain the accurate hand posture region. Then the extracted posture images are described with HOG feature and recognized by the ELM classifier. Experimental results show that the proposed hand posture recognition method achieves a high recognition rate and real-time processing.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant (No. 61172161), the National Natural Science Foundation for Distinguished Young Scholars of China under Grant (No. 61325007).

## References

1. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. *Artifical Intelligence Review*, vol. 38, pp. 1–54 (2012)
2. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and reviews*, vol. 37, no. 3, pp. 311–324 (2007)
3. Elmezain, M., Al-Hamadi, A., Michaelis, B.: Hand trajectory-based gesture spotting and recognition using HMM. In Proc. IEEE International Conference on Image Processing, pp. 3577–3580 (2009)
4. Weng, C., Li, Y., Zhang, M., Guo, K., Tang, X., Pan, Z.: Robust hand posture recognition integrating multi-cue hand tracking. In Proc. International Conference on E-learning and Games, Edutainment, pp. 497–508 (2010)
5. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-based 3D hand pose estimation from monocular video. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1793–1805 (2011)
6. Bagdanov, A.D., Del Bimbo, A., Seidenari, L., Usai, L.: Real-time hand status recognition from RGB-D imagery. In Proc. International Conference on Pattern Recognition, pp. 2345–2459 (2012)
7. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from sigle depth images. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
8. Argyros, A.A., Lourakis, M.I.A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In Proc. European Conference on Computer Vision, pp. 368–379 (2004)

9. Chen, L., Lin, H., Li, S.: Depth image enhancement for Kinect using region growing and bilateral filter. In Proc. IEEE Conference on Pattern Recognition, pp. 3070–3073 (2012)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
11. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing, vol. 70, pp. 489–501 (2006)
12. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics, vol. 2, no. 2, pp. 107–122 (2011)

# Real-Time Human Detection Based on Optimized Integrated Channel Features

Jifeng Shen<sup>1</sup>, Xin Zuo<sup>2</sup>, Wankou Yang<sup>3</sup>, and Guohai Liu<sup>1</sup>

<sup>1</sup> School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013, China

<sup>2</sup> School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, 212003, China

<sup>3</sup> School of Automation, Southeast University, Nanjing, Jiangsu, 210096, China

**Abstract.** We propose an optimized integrated channel features which can effectively improve the detection performance at the frame rate of 30 fps on images size of 640x480. The proposed method utilizes the distribution of filter response from positive and negative features to formulate the optimized combination of multiple filters. The optimized combination coefficient is learned from linear discriminative criterion which is superior to integrated channel features with random coefficients. Experimental results based on INRIA dataset shows the superiority of our method to other state-of-arts methods.

**Keywords:** Human detection, integrated channel feature, Adaboost.

## 1 Introduction

Real-time human detection is a very hot topic in computer vision field and attracts many researchers' attention in recent years. It is widely used in visual surveillance, behavior analysis or automated personal assistance field. Human detection in outdoor scene is a challenge task, which faces the difficulty of human deformation, illumination change, occlusion and scale transformation.

The notable landmarks of human detection is Dalas & Triggs, who presented a human detection method which made use of the Histogram of Orientation features[1] to model human silhouette and apply linear SVM to classify the stacked 3780 dimensional features generated by dense sampling. This method is very effective in detecting upright fully visible humans and robust to slight deformation. But the evaluation speed is more than 500ms to scan a 240x320 image that only have 1000 detection windows. This baffled its application in many of the field which needs real-time running speed. There're many papers[2-9] appeared which improve the HOG features and still cannot meet the requirement of real-time application. This survey[10] gives more detail of comparisons between different algorithms in this field. Most recently, Piotr[11] proposed integrated channel features which is easy to compute, but still face the bottleneck of overload of computation multiple scales of image features and image resizing, but it can be solved by the multi-scale feature approximation[12] which achieves the goal of real-time human detection in the application level. Although this

method runs very fast in real application, it still can be sped up with optimized ICF features, which is based on the statistic of distribution of filter response of positive and negative features.

In this paper, we propose the optimized ICF features, which learns the optimized coefficient of different feature channels and improved the discriminative ability. We make use of the multi-scale classifiers' framework, which greatly improve the detection rate and also reduce the running time.

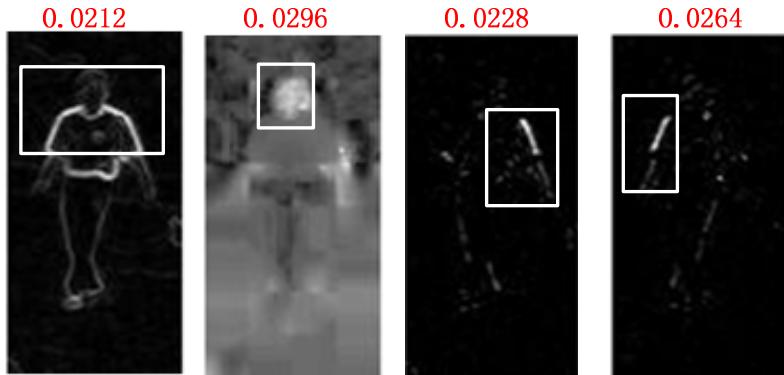
The rest of the paper is organized as follows: section 2 discuss the related work which our work based on. section 3 introduce our proposed OICF features. section 4 covers the experiment results. section 5 concludes the paper.

## 2 Integrated Channel Features (ICF)

Piotr first proposed the ICF feature [11], it models the feature  $C$  of image  $I$  as a channel generation function  $\Omega$ , so the feature of image  $I$  can be represent as  $C_i = \Omega_i(I)$ , where  $\Omega = \{\Omega_i\}, i=1,2,\dots,n$ .  $C_i$  is the ith feature for image I,  $\Omega_i$  is the ith channel generation function. The channel generation function can be linear, such as gray-level image of original image I or nonlinear, such as gradient image. Each channel represents a different feature space which derived from original image. The common feature channels can be gray image, different channel of RGB images and gradient image. Fig. 1 shows the different channels of the original image. We can see that different channels can reflect different aspect of input image. For example, the gradient image of different angle can reflect the lines which is similar to the Gabor filter, the canny image can represent the edge of human and different color space can reflect the color consistency in clothes.

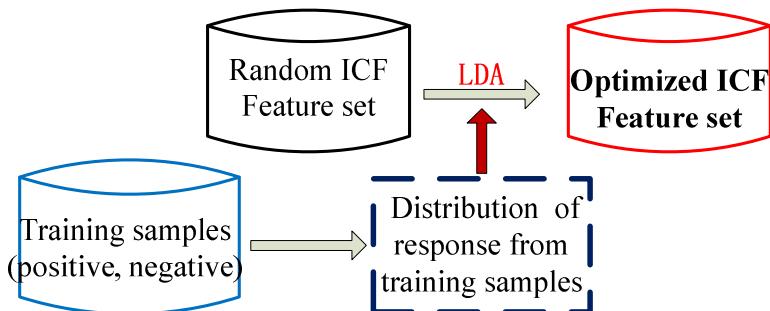


**Fig. 1.** Different channel of the input image  
(first row corresponds to the original, magnitude, gradient angle(1-6), sobel image;  
Second row corresponds to the LBP, orientation, LUV, HSV and canny image )



**Fig. 2.** Integral channel features  
(magnitude, chrome of LUV, third and fifth gradient images  
The decimal represent the random weight of each feature)

After build all the channels of input image, we can randomly sample in different channels to get informative area such that it can differentiate positive samples from negative samples. It can be formulated as follows. Supposing  $R_i(C)$  is the cumulative value in the ith rectangle area of channel  $C$ , the ICF feature value can be represent as  $\sum_i^c w_i R_i(C_i)$ , where  $w_i$  is the weight for the ith feature. The original ICF feature, both  $w_i$  and  $R_i$  are randomly generated and make use of Adaboost algorithm to do feature selection, finally construct the classifier. In each round of Adaboost algorithm, it generates a large pool of rectangles, such as 30000 areas, and find the most discriminative one. Fig. 2 shows one integral features from selected four channels of image I.



**Fig. 3.** Optimized ICF features

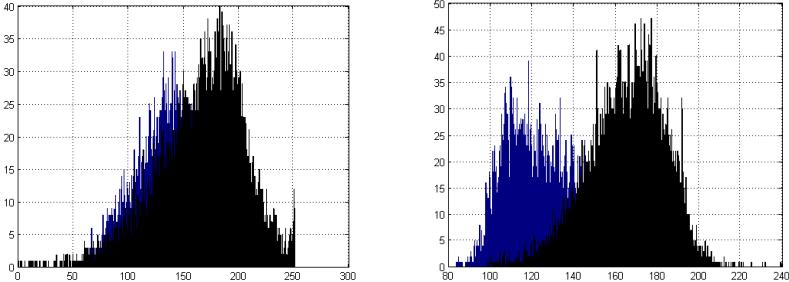
### 3 Optimized ICF Features

The original ICF feature suffers with the randomized feature combination, so the optimized performance of the detector cannot be guaranteed. We propose an optimized ICF features (OICF) which utilize the distribution of the positive and negative samples. The procedure of creating OICF feature is shown in Fig. 3. The key difference between ICF and OICF is that OICF is much discriminative than original ICF features due to applying the statistic information from training dataset. The ICF feature is defined as  $\sum_i^c w_i R_i(C_i)$  which can be reformulated as  $W^T R(C)$ , where  $W = [w_1, w_2, \dots, w_c]$  and  $R(C) = [R_1(C_1), \dots, R_c(C_c)]$ . In the original ICF feature, weight vector W is obtained from random values from uniform distribution and is not optimal for feature. Our proposed OICF feature learns the weight vector from positive and negative samples which can utilize the distribution of training data. The weight W can be learned by linear discriminative criterion which is shown in Eq.(1)

$$\mathbf{w}_{opt} = \arg \max_w \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \quad (1)$$

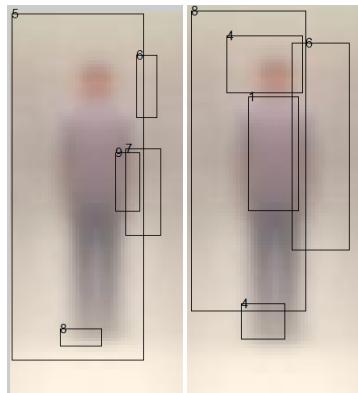
Where  $S_w$ ,  $S_b$  is the within-class scatter matrix and between-class scatter matrix and Eq.(1) can be solved with closed form.

In Fig. 4, we demonstrate the first selected histogram of original ICF feature and OICF feature of Adaboost algorithm for the 20000 positive and negative samples.



**Fig. 4.** First selected ICF feature and OICF feature for 20000 training data  
(Blue for positive samples, Black for negative samples)

In figure 5, we demonstrate the first selected ICF and OICF feature position, channel index and corresponding weights. In the ICF feature, it comprise of 5 channels, where the OICF feature also comprise of 5 channels. We can see that OICF feature located at the high frequency area of the image which can be more discriminative than the original ICF feature. the black rectangle indicate the area which our first weak classifier chosen, the number located at the top left of the rectangle represents the corresponding channel number.

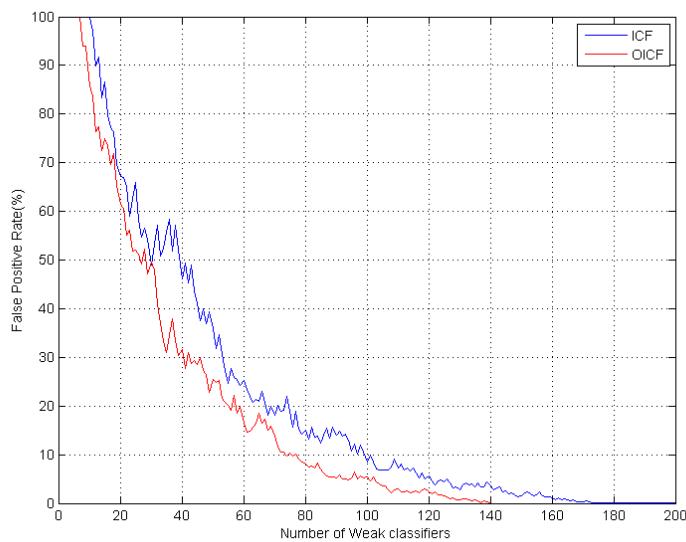


**Fig. 5.** Selected feature position for ICF and OCIF features

## 4 Experiment

### 4.1 Experiment Setting

In order to validate the effectiveness of our proposed features, we conduct experiment on INRIA database which is widely used in evaluating human detector. In training our human detector, 2416 human images and 2416 non-human images (randomly sampled from 1218 images) were used in the beginning, then bootstrap in the 1218 images in later training stage. All the training samples are cropped into 128x64. The final classifier is trained with soft cascade comprise of 2000weak classifiers.



**Fig. 6.** False positive rate comparison between ICF and OICF

## 4.2 Comparison between ICF and OICF

In order to compare the discriminative ability between ICF and our proposed OICF features, we have shown the curve of miss rate with number of weak classifiers which is shown in Fig. 6. From Fig. 6, we can see that, our proposed method is much faster to converge than original ICF features. our proposed method need only approximately 140 weak classifiers to get the miss rate of 0 with the true positive rate of 100% which is about 5% higher than ICF features. Furthermore, there is a large gap(20% lower miss rate with 40 weak classifiers) between two curves which indicates the effectiveness of our proposed method.

## 4.3 Comparisons with Other State-of-Art Algorithm

In order to validate the effectiveness of our OICF feature, we also compare our methods with other state-of-art algorithms with FPPI-miss rate curve, the result is shown in Fig. 7. From Fig. 7 we can see that our proposed method can gain the lowest miss rate at all given FPPI values, and it is 2 percent lower than the current state-of-art FPDW algorithm in average. It's worth to mention that we also implement our ICF feature which cannot reproduce the published result[11] and is still 1% higher than the original papers.

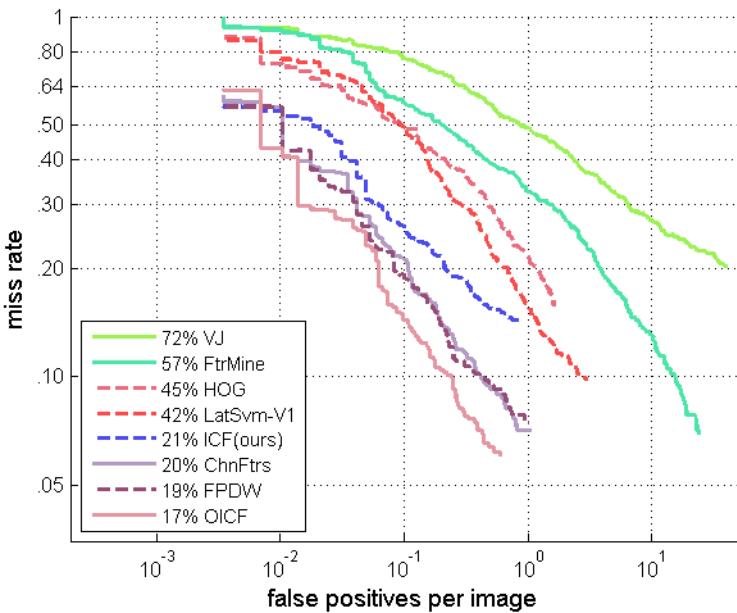


Fig. 7. FPPI Vs Miss rate

## 4.4 Samples Detection Result in INRIA Dataset

In Fig. 8, we demonstrate some of the detection results in INRIA dataset. The image in the upper rows are correctly detected all the true positives, where in the lower left and right image there is one false positives and 3 false negatives respectively. we can

see that the false positives actually is very similar to humans and the false negatives are caused by server occlusion and human attachment.



**Fig. 8.** Sample detection results



Fig. 8. (Continued)

#### 4.5 Runtime Comparisons

In this section, we will compare our proposed method with other state-of-arts methods. The results are shown in Tab. 1. The experiment is done on HP workstation ZBook 17(8 core CPU I7-4700MQ, 2.4GHZ, 32G) with Matlab 2013b. From Tab.1, we can see that both of our improved ICF and OICF feature can significantly improved the running speed, where OICF is fastest among all the methods.

**Table 1.**

Method	640x480 (fps)
ICF(ours)	25
ChnFtrs[11]	5
FPDW[12]	5
OICF(proposed)	<b>30</b>

#### 5 Conclusions

In this paper, we proposed a novel OICF features which utilizing the distribution of filter response from multiple channels of positive and negative samples. the proposed feature can obtain great efficiency improvement compared with state-of-art integrate channel features. the results shows the importance of choosing optimized combination coefficient of multiple channels. experimental results indicate that by using OCIF features, both the speed and accuracy of detector can be improved.

**Acknowledgement.** This project is supported by NSF of China(61005008), Nature Science Foundation of the Jiangsu Higher Education Institutes of China (12KJB520003), Young Scientist Foundation of Jiangsu Province(BK20140566).

#### References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
2. Sabzmeydani, P., Mori, G.: Detecting Pedestrians by Learning Shapelet Features. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 90–97 (2007)
3. Wang, X., Han, T., Yan, S.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: IEEE International Conference on Computer Vision (2009)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
5. Lin, Z., Davis, L.S.: A pose-invariant descriptor for human detection and segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 423–436. Springer, Heidelberg (2008)
6. Chen, Y., Chen, C.: Fast Human Detection Using a Novel Boosted Cascading Structure With Meta Stages. IEEE Transactions on Image Processing 17(8), 1452–1464 (2008)

7. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
8. Walk, S., Majer, N., Schindler, K., Schiele, B.: New Features and Insights for Pedestrian Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
9. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human Detection Using Partial Least Squares Analysis. In: ICCV2009, pp. 24–31 (2009)
10. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4), 743–761 (2012)
11. Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral Channel Features. In: British Machine Vision Conference 2009, London, England (2009)
12. Dollar, P., Belongie, S., Perona, P.: The Fastest Pedestrian Detector in the West. In: BMVC 2010 (2010)

# Facial Feature Extraction Based on Robust PCA and Histogram

Xiao Luan and Weisheng Li

Chongqing Key Laboratory of Computational Intelligence,  
Chongqing University of Posts and Telecommunications, Chongqing 400065, China  
[{luanxiao, liws}@cqupt.edu.cn](mailto:{luanxiao, liws}@cqupt.edu.cn)

**Abstract.** Inspired by recently-proposed robust principal component analysis (RPCA), in this paper we present a feature extraction method for robust face recognition in the presence of random pixel corruption and occlusion. Unlike most work focusing on the low-rank structure recovered by RPCA, we consider that the sparse error component contains more discriminating power which is essential to face recognition. In order to illustrate the intensity distribution of the sparse error component, a histogram-based sparsity measure is introduced for feature extraction. Compared with the related state-of-the-art methods, experimental results on Extended Yale B database verify the advancement of the proposed method for partially corrupted and occluded face images.

**Keywords:** Face recognition, Robust PCA, Sparse, Histogram.

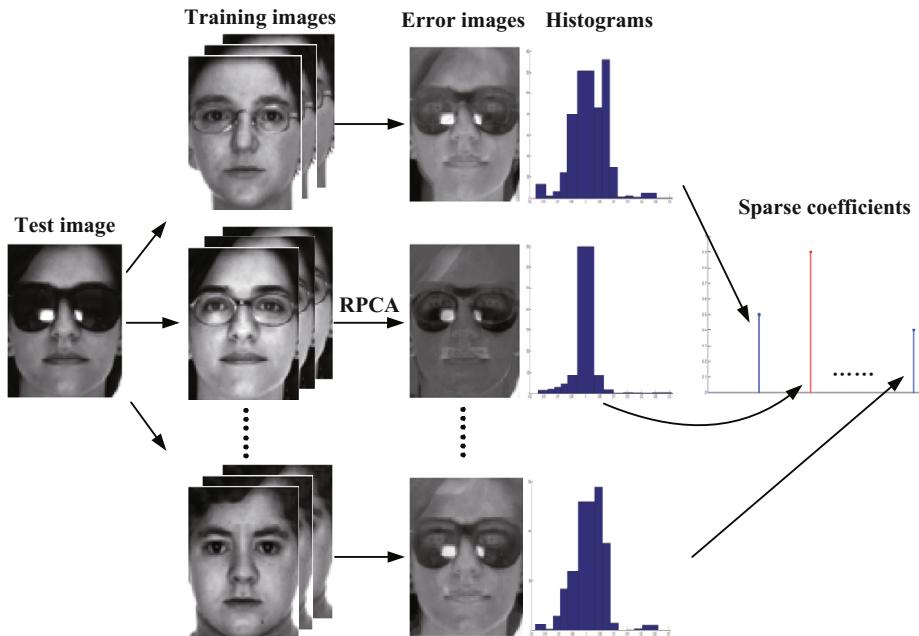
## 1 Introduction

Face recognition has been an active topic of biometrics for both its scientific challenges and its wide potential applications[6]. Most of the systems to date can only successfully recognize faces when images are acquired under constrained conditions. However, their performance will degrade abruptly when facial images are captured under varying illuminations, poses, and expressions, especially for the case of pixel corruption or partial occlusion [17,1,11]. Therefore, recognition with corruption or incomplete data is a great challenge.

In terms of feature extraction, face recognition approaches are broadly classified into two categories, i.e., holistic based approaches and local approaches. By extracting holistic face features, conventional holistic approaches such as PCA [13], LDA [3] aim to find projections that can well separate the classes in lower dimensional spaces. These approaches are known to yield better results on several public face databases. Nevertheless, the performance of these approaches will be seriously degraded under partially occluded faces. Local features [9,16,2,7], which are computed from a small fraction of the whole image pixels, are less likely to be corrupted by occlusion than holistic features. Thus local approaches try to extract meaningful partial facial features that can eliminate or compensate the difficulties brought by occlusion variations [15]. Recently, face recognition

approaches based on linear representations have led to the state-of-the-art performance. The most representative approaches are sparse representation-based classification (SRC) [15] and linear regression-based classification (LRC) [12]. SRC aims to find a sparse representation of the query image in terms of an over-complete basis, in addition to a sparse error corresponding to the occlusion component. LRC represents a test image as a linear combination of training images of class-specific samples. Experimental results on several publicly databases have shown the efficacy of those two algorithms, while their performance are not satisfactory with heavily corrupted data.

Recently a sparse learning framework named Robust Principal Component Analysis (RPCA) [4,14] is proposed. Given an observed data matrix which is the sum of a low-rank component and a sparse component, RPCA can exactly recover each component, even though a positive fraction of its entries are arbitrarily corrupted. As a powerful algorithm for data analysis, RPCA draws much attention from many research areas.



**Fig. 1.** Overview of our approach. Our method uses RPCA to decompose test image into sparse error image with respect to each individual (middle), and computes the sparseness of intensity histogram of those error images. Red coefficients correspond to the correct individual.

In this paper, we attempt to address the problem of face recognition under pixel corruption or occlusion from a perspective of feature extraction. Based on

our previous work [10] that representing characteristic of sparse error component from intensity domain and gradient domain, we represent these facial features in a more intuitive way by introducing a histogram-based sparsity measure, and the true individual can be identified from the sparse coefficients, as shown in Fig. 1. Note that in our previous work, we need to use a parameter balancing two descriptors (i.e., sparsity and smoothness), while in this paper we show that if sparsity in feature extraction is properly harnessed, one descriptor is enough for our method to make a satisfactory performance.

Section 2 briefly reviews RPCA. Section 3 elaborates on the proposed algorithm. Section 4 verifies the proposed approach with experiments on popular face database, comparing with several state-of-the-art face recognition techniques. The conclusion is drawn in Section 5.

## 2 Robust Principal Component Analysis (RPCA)

Assume a given large data matrix  $D \in \mathbb{R}^{m \times n}$  has a low-rank structure  $L$  yet corrupted by sparse errors component  $E$ , i.e.,  $D = L + E$ . The objective is to reliably recover the low-rank component  $L$  from the highly corrupted matrix  $D$ . The initial formulation of Robust Principal Component Analysis (RPCA) [14] can be described as follows:

$$\min_{L, E} (\text{rank}(L) + \gamma \|E\|_0), \quad \text{s.t. } D = L + E \quad (1)$$

where  $\|E\|_0$  denotes the matrix  $\ell_0$ -norm, i.e., counting nonzero elements in the matrix  $E$ . However, (1) is difficult to solve due to the non-convexity and non-smoothness of the rank measure and zero-norm penalty. Through tractable convex optimization, Candès et al. [4] solve Principal Component Pursuit (PCP) in the following relaxed form:

$$\min_{L, E} (\|L\|_* + \gamma \|E\|_1), \quad \text{s.t. } D = L + E \quad (2)$$

where the rank operation in (1) is replaced by matrix nuclear norm  $\|\cdot\|_*$  (i.e., the sum of the singular values), the matrix  $\ell_1$ -norm (i.e., the sum of absolute matrix entries) approximates the matrix  $\ell_0$ -norm and,  $\gamma$  is the regularization parameter for balancing. It has been shown both theoretically and empirically that, under rather weak assumptions, the solution of (2) perfectly recovers the low-rank and the sparse component, as long as the rank of  $L$  is not too large and the errors  $E$  is sparsely supported [4].

## 3 Proposed Method

In the context of face recognition, we identify a  $a \times b$  gray scale image with the vector  $\nu \in \mathbb{R}^m$  ( $m = ab$ ) given by stacking its columns. Let  $n_i$  be the training images of each subject  $i$  stacking as columns of a matrix  $D_i \in \mathbb{R}^{m \times n_i}$ . Denote by  $D = [D_1, D_2, \dots, D_K] \in \mathbb{R}^{m \times n}$  the training images of all  $K$  subjects. For

different kinds of face images, we can use RPCA to decompose these images into low-rank components and sparse error components. The low-rank components stand for common structures or intrinsic ingredients, while the sparse error components represent different facial variations i.e., illumination, expression and corruption. Furthermore, given a test face image under different subjects, it has been revealed that it is the sparse error images instead of low-rank components contain discriminative nature which is beneficial to face classification [10]. Next, we will show how to illustrate the differences between those error images.

Given a test image  $y \in \mathbb{R}^m$ , we first perform RPCA to obtain the error images denoted by  $E_i$ ,  $i = 1, 2, \dots, K$ . In our previous work, based on an intuitive observation, we simply count the number of elements within a relatively small range in sparse error image as the defined sparsity descriptor [10]. However, only exploiting sparsity from the pixel domain to describe the error image's difference is not enough, so another descriptor named smooth descriptor is defined from the gradient domain. Finally, a weighted based method that uses these two descriptors jointly to extract the facial features is proposed, while a parameter  $\alpha$  balancing the weight between sparse and smooth need to set in practice. Hence, we hope to design an algorithm does not need to deal with model parameters, which makes it very simple to use in practice.

In statistics, histogram is a graphical representation of data distribution, and is also an estimate of the probability distribution of a continuous variable. Through the histogram, we can get many features, e.g., spread (i.e., the scale) of the data, presence of outliers and presence of multiple modes in data. Therefore, we introduce histogram to represent the intensity distribution of sparse error images. For each histogram,  $x$ -axis stands for the interval of intensity and  $y$ -axis stands for the frequency of points falling into associated interval. If we stack the frequency as a vector  $E_i$ ,  $i = 1, 2, \dots, K$ , then we need to find a proper measure to quantitatively represent the feature of histogram vector.

In this paper, we adopt the sparseness measure [5] which is based on the relationship between the  $\ell_1$  norm and the  $\ell_2$  norm:

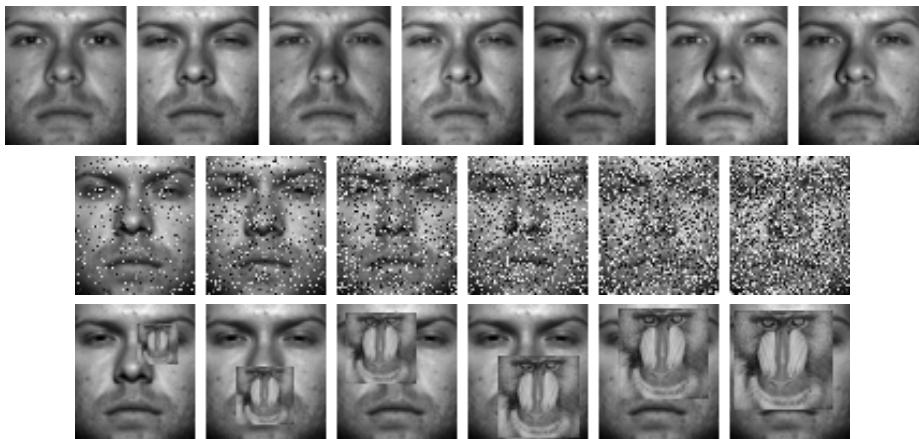
$$\text{sparseness}(\mathbf{x}) = \frac{\sqrt{n} - \sum |x_i| / \sqrt{\sum x_i^2}}{\sqrt{n} - 1} \quad (3)$$

where  $n$  is the dimensionality of vector  $\mathbf{x}$ . This function evaluates to unity if and only if  $\mathbf{x}$  contains only a single non-zero component, and takes a value of zero if and only if all components are equal (up to signs), interpolating smoothly between the two extremes. For our method, we let  $\mathbf{x} = E_i$ . Finally, the decision rules in favor of the class with larger sparseness. As shown in Fig. 1, by the use of histogram representation of these sparse error images, the data distribution of error image from the true individual (marked in red line) clearly contain discriminative information than incorrect individuals.

## 4 Experimental Results

Experiments are conducted on the Extended Yale B database [8] for face recognition, which consists of 2,414 frontal face images of 38 individuals. All test image

data are cropped and re-sized to  $64 \times 56$  images. The database is divided into five Subsets according to different lighting conditions. For the following experiments Subset 1 is selected as training samples, and Subset 2 for testing samples, respectively.



**Fig. 2.** Samples from Extended Yale B database. Top row: training samples from Subset 1. Middle row: Test Images of Subset 2 under varying levels of pixel corruption, from 10% to 60%. Bottom row: Test Images of Subset 2 under varying levels of block occlusion, from 10% to 60%.

**Recognition with Random Pixel Corruption.** We first examine the ability of the proposed method on dealing with random pixel corruption, comparing to several state-of-the-art techniques. For each testing image, we replace a certain percentage of its pixels by uniformly distributed random values within  $[0, 255]$ . The corrupted pixels are randomly chosen for each test image and the locations are unknown to the algorithm. The middle row of Fig. 2 shows a test images from Subset 2 with 10 to 60 percent pixel corruption. Table 1 lists the comparison results of those method.

**Table 1.** Recognition rate (%) under varying levels of pixel corruption

**Recognition Despite Random Block Occlusion.** We simulate various levels of contiguous occlusion, from 10 percent to 60 percent, by replacing a square block of each test image with a baboon image. The location of occlusion is randomly chosen for test image and is unknown to the algorithms. An example of test images from Subset 2 with 10 to 60 percent occlusion is shown in the bottom row of Fig. 2, and the recognition results are shown in Table 2. As shown in Table 2, the proposed method correctly classifies all subjects.

**Table 2.** Recognition rate (%) under varying levels of block occlusion

Percent occluded	10%	20%	30%	40%	50%	60%
Eigenfaces + NN	85.53	81.58	75.22	57.89	44.3	25.66
Fisheerfaces + NN	100	99.56	81.8	45.61	53.73	28.51
LRC	99.34	98.03	94.74	72.15	49.12	22.59
SRC	100	98.9	94.96	67.54	41.45	17.54
<b>proposed method</b>	<b>99.78</b>	<b>100</b>	<b>100</b>	<b>95.61</b>	<b>89.04</b>	<b>78.07</b>

## 5 Conclusion

We present a feature extraction method to recognize face images under pixel corruption or occlusion. By introducing a histogram-based sparsity measure, the true individual can be identified from the sparse coefficients. Our method does not need to deal with model parameters, which makes it very simple to use in practice. Comparison with the state-of-the-art algorithms show the efficacy of the proposed method.

**Acknowledgments.** This work is supported by the Program for National Natural Science Foundation of China (No. 61272195) and New Century Excellent Talents in University of China(NCET-11-1085).

## References

1. Abate, A., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D face recognition: A survey. *Pattern Recognition Letters* 28(14), 1885–1906 (2007)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12), 2037–2041 (2006)
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 711–720 (1997)
4. Candes, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *Journal of the ACM* 58(3), 1–11 (2011)
5. Hoyer, P., Dayan, P.: Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* 5, 1457–1469 (2004)

6. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14(1), 4–20 (2004)
7. Jia, H., Martinez, A.: Face recognition with occlusions in the training and testing sets. In: *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–6 (2008)
8. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(5), 684–698 (2005)
9. Li, S., Hou, X., Zhang, H., Cheng, Q.: Learning spatially localized, parts-based representation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 207–212 (2001)
10. Luan, X., Fang, B., Liu, L., Yang, W., Qian, J.: Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion. *Pattern Recognition* 47(2), 495–508 (2014)
11. Luan, X., Fang, B., Liu, L., Zhou, L.: Face recognition with contiguous occlusion using linear regression and level set method. *Neurocomputing* 122(25), 386–397 (2013)
12. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(11), 2106–2112 (2010)
13. Turk, M., Pentland, A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
14. Wright, J., Ganesh, A., Rao, S., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: *Neural Information Processing Systems* (2009)
15. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)
16. Yuen, P.C., Lai, J.H.: Face representation using independent component analysis. *Pattern Recognition* 35(6), 1247–1257 (2002)
17. Zhao, W., Chellappa, R.: *Face Processing: Advanced modeling and methods*. Academic Press (2006)

# Multimodal Finger Feature Fusion and Recognition Based on Delaunay Triangular Granulation

Jinjin Peng, Yanan Li, Ruimei Li, Guimin Jia, and Jinfeng Yang

Tianjin Key Lab for Advanced Signal Processing  
Civil Aviation University of China, Tianjin, China  
jfyang@cauc.edu.cn

**Abstract.** For personal identification, three modalities of fingers, fingerprint (FP), finger-vein (FV) and finger-knuckle-print (FKP), can be used respectively. Fusing these modalities together as a whole biometric measure should naturally highlight the finger superiority in convenience and universality as well as recognition accuracy improvement. In this paper, a new finger recognition method based on granular computing is proposed. This method can synergistically combine the features of FP, FV and FKP in feature level and provide robustness to finger pose variation. The proposed granular space is constructed in bottom-up manner with three granule-layers. And a coarse-to-fine scheme is used for granule matching. Experiments are performed on a self-built database with three modalities to validate the proposed method in personal identification.

**Keywords:** Multimodal biometrics, Finger-knuckle-print, Finger-vein, Finger-print, Granular computing.

## 1 Introduction

As the traditional identifiers are easy to be forgotten, stolen or misplaced, biometric based identification techniques have developed rapidly with the increasing security demand [1]. Most biometric systems used in practical applications are unimodal, and they often confront with some problems, such as noisy data, intra-class variations, inter-class similarities, and spoof attacks. To address some limitations imposed by unimodal biometric, multimodal biometric-based technology is proved to be an effective method [1,2].

There are four levels of fusion in a multimodal biometric system: pixel level, feature level, matching score level and decision level [2]. Among these fusion levels, feature level fusion has been believed that it could provide the best recognition results [3]. Nevertheless, fusion at this level is difficult to achieve because multimodal spaces may be incompatible in many cases. At present, some attempts have been done toward the feature level fusion [4,5]. However, there still two problems, feature vector with a high dimensionality and huge computational cost. Furthermore, some scholars used canonical correlation analysis (CCA) method for feature level fusion [6], but CCA cannot work well beyond three modalities.

Granular computing (GrC) is a new information processing concept and computing paradigm [7,8]. GrC is a human-centered approach that can solve problems using knowledge from multiple levels of information granularity. Since Zadeh introduced

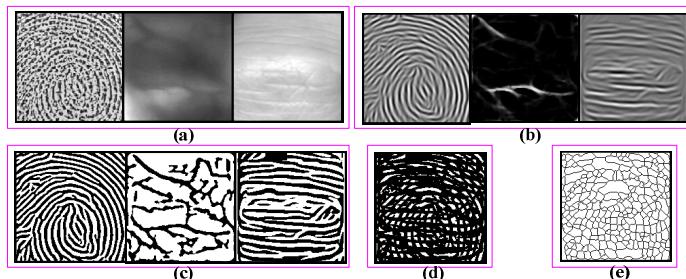
and discussed the notion of information granulation in 1979 [9], GrC has been rapidly developed and implemented in information processing [10,11,12]. Recently, Zheng proposed tolerance granular space model and applied it to texture recognition [13], Chan used granulation method for pedestrian detection [14], Bhatt applied granular feature to face recognition [15]. These researches indicate that image analysis and recognition is a new stage for GrC in computer vision.

In this paper, we use GrC to solve a finger-based recognition problem. Here, a finger trait is composed of FP, FV and FKP. A three-layer granular model is developed to solve the fusion recognition problem of FP, FV and FKP. To address the consistency and compatibility of FP, FV and FKP, we normalize the three modalities with same aspect. The construction of the basic granules is a key issue in granular process, therefore, we use Delaunay triangulation based on intersection feature points to obtain basic granules, whose intension and extension is respectively Gabor feature and triangle. Then, we can construct a bottom-up multi-granularity granular model based on the basic granules [13]. The recognition approach is a top-down method. To evaluate the performance of this method, a self-built database with three modalities is used here. The experimental results validate that our method performs reliable and precise in feature fusion and personal identification.

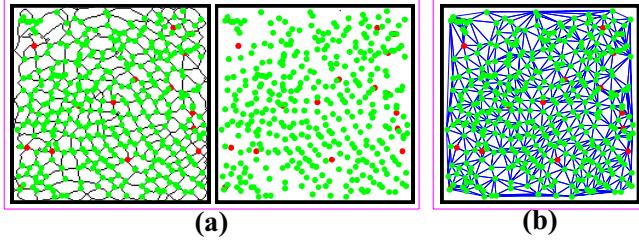
## 2 Triangulation

The traditional granulation methods often divide the granules in space without considering the spatial structure of the original object. Here, we introduce the topological structure of intersection feature points to construct the basic granules.

In order to improve the compatibility of the three modalities, the images of FP, FV and FKP have been normalized to the same size  $170 \times 170$ , as shown in Fig. 1. (a) (Left: FP, Middle: FV, Right: FKP). As the quality of the acquired images is generally poor, we need to enhance the ROI images to strengthen the blurred information. Due to the different imaging principle and type of texture of the three modalities, we enhance FV by Gabor filter [16] and conduct Steerable filter on FP and FKP [17]. The results are provided in Fig. 1. (b) (Left: FP, Middle: FV, Right: FKP).



**Fig. 1.** The preprocessing results. (a) Normalization images. (b) Filtered images. (c) Binary images. (d) Combined binary image. (e) Skeleton of (d).



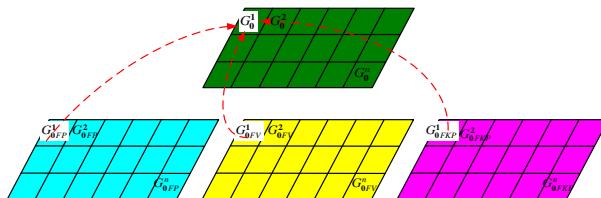
**Fig. 2.** Intersection feature points extraction and triangulation. (a) Intersection feature points extraction. (b) Triangulation.

Then, a threshold method is used to obtain the binary images of the three modalities [18], the results are shown in Fig. 1. (c) (Left: FP, Middle: FV, Right: FKP). In order to provide enough intersection feature points and ensure that the intersection feature points in the three modalities are same, we superimpose the binary images of the three modalities, and obtain a fusion binary image, as show in Fig. 1. (d). Then, we extract the skeleton of the fusion binary image use the thinning algorithm proposed in [18]. The result is shown in Fig. 1. (e).

As the fusion image is the combination of binary images, there are a lot of intersection feature points in the skeleton image. To obtain the intersection feature points, a pixel-wise operation for a  $3 \times 3$  region proposed in [19] is used here. The results are shown in Fig. 2. (a). Based on the detected intersection feature points, we can obtain stable and unique triangles using Delaunay triangulation. The result is shown in Fig. 2. (b).

### 3 Granular Initialization

We use a 2-tuple  $G = (IG, EG)$  to represent granule.  $IG$  is the intension of a granule, which describes the attribute of the granule.  $EG$  is the extension of the granule, which is a set containing all the objects in the granule. In this paper, we use  $(i_1, i_2, \dots, i_n)$  to represent  $IG$ , and use  $(e_1, e_2, \dots, e_m)$  to describe the geometric characteristic of  $EG$ .



**Fig. 3.** Basic granules fusion

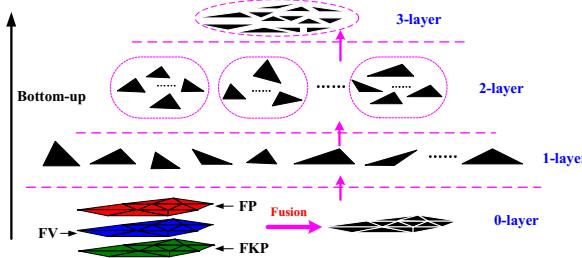
Each pixel in an image modality can be regarded as a granule  $G_{0\text{modal}}^i$ . The vector expression of  $IG_{0\text{modal}}^i$  consists of 8 Gabor coefficients obtained by Gabor transformation [19], that is,  $IG_{0\text{modal}}^i = (g_{0\text{modal}}^{i1}, g_{0\text{modal}}^{i2}, \dots, g_{0\text{modal}}^{i8})$ .  $EG_{0\text{modal}}^i$  is the  $i$ th pixels

in three image modalities. The granules with the same coordinate in FP FV and FKP images are fused to form original granules  $G_0^i$  ( $i = 1, \dots, n$ ),  $n$  is the number of the pixels in a finger image combined by three image modalities. Here,  $IG_0^i = (IG_{0FP}^i, IG_{0FV}^i, IG_{0FKP}^i)$ ,  $EG_0^i$  is the  $i$ th pixel in a finger image. The fusion process is shown in Fig. 3.

## 4 Granulation and Recognition

### 4.1 Three-Layer Granulation

As mentioned above, a normalized image contains 28900 pixels, so GrC process is very time consuming using the original pixel-based granules. To improve granular matching efficiency, we thus construct a hierarchical granular recognition model in a bottom-up manner, which has three granular layers as shown in Fig. 4. The granular process is described as follows.



**Fig. 4.** 3-layer bottom-up granulation process

**Basic Granule Generation.** Based on Delaunay triangulation in Section 2, the obtained triangles are regarded as basic granules in the first granular-layer. Let  $G_1^j$  represent a basic granule corresponding to the  $j$ th triangle,  $IG_1^j$  and  $EG_1^j$  respectively denote the intension and the extension of  $G_1^j$ .  $EG_1^j$  is a granule-set composed of the pixel-based granules in  $j$ th triangle.  $IG_1^j$  is defined as a feature vector that consists of local Gabor feature  $LG_1^j$  and triangle shape feature  $SG_1^j$ . Here,  $LG_1^j = (g_{1FP}^{j1}, \dots, g_{1FP}^{j8}, g_{1FV}^{j1}, \dots, g_{1FV}^{j8}, g_{1FKP}^{j1}, \dots, g_{1FKP}^{j8})$  is an average absolute deviation (AAD) Gabor feature of pixel-based granules in  $j$ th triangle.  $SG_1^j = (h_{1FP}^{j1}, \dots, h_{1FP}^{j7}, h_{1FV}^{j1}, \dots, h_{1FV}^{j7}, h_{1FKP}^{j1}, \dots, h_{1FKP}^{j7})$  is a Hu moment feature of this triangle. Since AADs may be same for two different triangles, as shown in Fig. 5, the introduction of Hu moments [21] can increase the discrimination of  $IG_1^j$ .



**Fig. 5.** Two schematic triangle granules

**Granule Clustering.** K-means analysis is used to construct the second granular-layer based on  $LG_1^j$ . The number of clusters K is obtained by DBE algorithm [21]. Each cluster can be viewed as a 2-layer granule.  $IG_2^k$  and  $EG_2^k$  respectively denote the intension and the extension of  $G_2^k$  ( $k = 1, \dots, K$ ).  $LG_2^k$  is an AAD Gabor feature of  $LG_1^j$ ,  $SG_2^k$  is Hu moments corresponding to the region constituted by the triangles in  $k$ th cluster.

**The 3-Layer Granule Construction.** A 3-layer granule is synthesized by all the 2-layer granules,  $LG_3$  is an AAD Gabor feature of  $LG_2^k$ ,  $SG_3$  is Hu moments corresponding to the convex polygon composed by all the triangles.  $EG_3$  contains all the 2-layer granules. Thus, each finger can be represented by a granule  $G_3$ .

## 4.2 Recognition

The top-down granular recognition process is consistent with the granular process. The local Gabor feature (LG) and shape feature (SG) of a granule are concatenated into a vector as a granular descriptor (GrD). Since the discriminabilities of intension in local Gabor features and shape features vary with granular layers. We define  $GrD_l = (w_l LG_l, v_l SG_l)$  ( $w_l + v_l = 1$ ) in the  $l$ th granular layer,  $w_l$  and  $v_l$  vary with granular layers. Here, we use the 3-layer granules and 2-layer granules for recognition. The recognition process is described as follows.

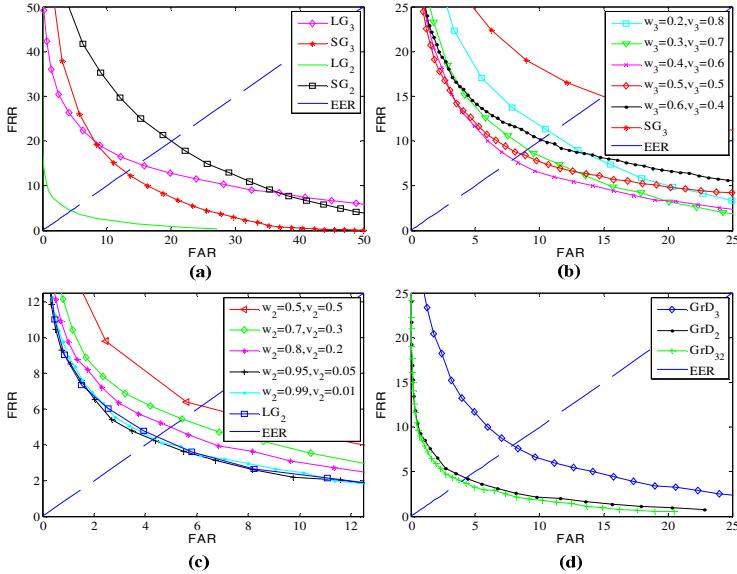
**Match in Third Layer.** For two finger images to be matched, we define the similarity between  $GrD_3$  and  $GrD'_3$  as  $Sim_3 = \cos(GrD_3, GrD'_3)$ . If  $Sim_3 \geq T_3$  ( $T_3$  is the decision threshold in the third layer), the two granules are similar, namely, the two finger images match in third layer. Because the 3-layer granules are extremely coarse, a successful matching behavior in the third granule layer cannot ensure that the two finger images are from the same individual. Then, we need further conduct granule matching in the second layer.

**Match in Second Layer.** Assume there are  $M_2$  2-layer granules in an input finger image, and  $N_2$  2-layer granules in a sample image. Calculate the similarity between  $GrD_2^m$  and  $GrD_2'^n$  by  $Sim_{2mn} = \cos(GrD_2^m, GrD_2'^n)$  ( $m = 1, 2, \dots, M_2$ ,  $n = 1, 2, \dots, N_2$ ). If  $Sim_{2mn} \geq T_2$  ( $T_2$  is the decision threshold in the second layer), the two granules are similar, we then calculate the match score of the two finger images in the second layer using  $Score_2 = 2S_2 / (M_2 + N_2)$ . If  $Score_2 \geq Ts_2$  ( $Ts_2$  is the matching threshold in the second layer), the two finger images match in the second layer. If both  $Sim_3 \geq T_3$  and  $Score_2 \geq Ts_2$  are satisfying, we think that the two finger images are from an identical individual.

## 5 Experiments

Here, a self-built database that contains 600 finger-vein images, 600 fingerprint images and 600 finger-knuckle-print images from 60 individuals is used in the experiments.

To prove the feasibility and effectiveness of our algorithm, we carry out several experiments. Here, we respectively use LG, SG and GrD for granule matching in different layers. The threshold  $Ts_2$  is an empirical value ( $Ts_2 = 0.5$  is using here).



**Fig. 6.** ROC curves (a) Using LG and SG alone (b) Comparison of using different weights in third layer (c) Comparison of using different weights in second layer (d) Comparison of single layer and multi-layer

Firstly, we solely use LG or SG as feature vector for matching. The results are shown in Fig. 6. (a). Then, we use GrD with different  $w_l$  and  $v_l$  for granule matting, the results are shown in Fig. 6. (b) and Fig. 6. (c). Fig. 6. (a) shows that LG and SG have different abilities in distinguishing granules in different layers. Fig. 6. (b) and Fig. 6. (c) illustrate that LG can express image characteristics more effectively in low layer than in high layer, while SG is inverse. The reason is that the fine granules of different individuals may have similar shapes, but their local features are significantly different, and only when they are combined together, they are able to distinguish the granules between two different individuals effectively. Fig. 6. (d) reports the recognition results using GrD with weights in different layers and a coarse-to-fine multi-layer matching approach mentioned in Section 4. The optimal weights are  $w_2 = 0.95$ ,  $v_2 = 0.05$  and  $w_3 = 0.3$ ,  $v_3 = 0.7$  here. In multilayer-granular recognition, the principle of choosing the threshold is that ensuring that false rejection rate (FRR) is zero and false acceptance rate (FRR) is as low as possible in high layer according to empirical knowledge. The

experiments (see Fig. 6. (d)) show that the recognition results using second layer granules are better than those using third layer granules when both considering local features and shape features. Fig. 6. (a) also shows that the multilayer-granular recognition accuracy is the best. The reason is that the discrimination is enhanced using multilayer granular information. Besides, multilayer-granule recognition also can reduce the time cost since the non-matched granules in the third layer are neglected for granule matching in the second layer.

## 6 Conclusion and Future Work

In this paper, a new biological recognition method based on three finger traits (FP, FV and FKP) was proposed. We combined granular computing with multimodal biometric fusion identification, and constructed a bottom-up 3-layer granular model. Base on the same database, we several experiments were implemented. The results showed that combining the local feature and shape feature of a granule could obtain higher recognition accuracy than using a single characteristic of the granule. Moreover, a coarse-to-fine matching scheme was used, which could save computational cost. However, the proposed method still has much room for improvement in constructing basic granules and feature extraction.

**Acknowledgements.** This work is jointly supported by National Natural Science Foundation of China (No.61379102) and The Fundamental Research Funds for the Central Universities (No. 3122014C003).

## References

1. Ross, A., Jain, A.K.: Multimodal biometrics: an overview. In: XII European Signal Processing Conference, vol. 2, pp. 1221–1224 (2004)
2. Sim, T., Zhang, S.: Continuous Verification Using Multimodal Biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 687–700 (2007)
3. Gudavalli, M., Babu, A.V., Raju, S.V., Kumar, D.S.: Multimodal Biometrics—sources, Architecture & Fusion Techniques: An Overview. In: International Symposium on Biometrics and Security Technologies, pp. 27–34. IEEE Press, New York (2012)
4. Yang, J., Zhang, X.: Feature-level fusion of fingerprint and finger-vein for personal identification. Pattern Recognition Letters 33(5), 623–628 (2012)
5. Gawande, U., Zaveri, M., Kapur, A.: Bimodal biometric system: feature level fusion of iris and fingerprint. Biometric Technology Today 2013(2), 7–8 (2013)
6. Schreier, P.J.: A Unifying Discussion of Correlation Analysis for Complex Random Vectors. IEEE Transactions on Signal Processing 56(4), 1372–1336 (2008)
7. Miao, D., Wang, G., Liu, Q., Lin, T.Y., Yao, Y.: Granular Computing: Past, Present and Future Prospects. Science Press (2007) (in Chinese)
8. Yao, Y.: Interpreting Concept Learning in Cognitive Informatics and Granular Computing. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 39(4), 855–866 (2009)
9. Zadeh, L.A.: Fuzzy Sets and Information Granulation Advances in Fuzzy Set Theory and Application. North Holland Publishing Press, Netherlands (1979)

10. Butenkov, S.A.: Granular Computing in Image Processing and Understanding. In: IASTED International Conference on Artificial Intelligence and Application, vol. 2, pp. 811–816 (2004)
11. Bargiela, A., Pedrycz, W.: Toward a Theory of Granular Computing for Human-Centered Information Processing. *IEEE Transactions on Fuzzy Systems* 16(2), 320–330 (2008)
12. Yao, J., Vasilakos, A.V., Pedrycz, W.: Granular Computing: Perspectives and Challenges. *IEEE Transactions On Cybernetics* 43(6), 1977–1989 (2013)
13. Zheng, Z.: Image Texture Recognition based on Tolerance Granular Space. *Journal of Chongqing University of Posts and Telecommunications* 21(4), 484–489 (2009) (in Chinese)
14. Chan, Y., Fu, L., Hsiao, P., Lo, M.: Pedestrian Detection Using Histograms of Oriented Gradients of Granule Feature. In: 4th IEEE Intelligent Vehicles Symposium, pp. 1410–1415. IEEE Press (2013)
15. Bhatt, H.S., Bharadwaj, S., Singh, R., Vatsa, M.: Recognizing Surgically Altered Face Images using Multi-objective Evolutionary Algorithm. *IEEE Transactions on Information Forensics and Security* 8, 89–100 (2013)
16. Yang, J., Yang, J.: Multi-Channel Gabor Filter Design for Finger-vein Image Enhancement. In: 5th International Conference on Image and Graphics, pp. 87–91. IEEE Press, New York (2009)
17. Freeman, W.T., Adelson, E.H.: The Design and Use of Steerable Filter. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 13, 891–906 (1991)
18. Vlachos, M., Dermatas, E.: Vein segmentation in infrared images using compound enhancing and crisp clustering. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 393–402. Springer, Heidelberg (2008)
19. Yu, C., Qin, H., Zhang, L., Cui, Y.: Finger-Vein Image Recognition Combining Modified Hausdorff Distance with Minutiae Feature Matching. *Journal of Biomedical Science and Engineering* 1(4), 280–289 (2009)
20. Hu, M.K.: Visual Pattern Recognition by Moment Invariants. *IEEE Transaction on Information Theory* 8(2), 179–187 (1962)
21. Wang, L., Leckie, C., Ramamohanarao, K., et al.: Automatically Determining the Number of Clusters in Unlabeled Data Sets. *IEEE Transactions on Knowledge and Data Engineering* 21(3), 335–350 (2009)

# Robust Face Recognition via Facial Disguise Learning

Meng Yang and Linlin Shen

Shenzhen Key Laboratory of Spatial Smart Sensing and Services  
School of Computer Science and Software Engineering, Shenzhen University

**Abstract.** The sparse representation based classifier (SRC) has been successfully applied to robust face recognition (FR) with various disguises. Following SRC, recently regularized robust coding (RRC) was proposed for more robustness to facial occlusion by designing a new robust representation residual term. Although RRC has achieved the leading performance, it ignores the prior knowledge embedded in facial disguises. In this paper, we proposed a novel facial disguise learning (FDL) model, in which the unknown occlusion pattern in the query image is learned using a collected disguise mask dictionary. Two learning strategies with an iterative reweighted coding algorithm, independent FDL and joint FDL, were presented in this paper. The experiments on face recognition with disguise clearly show the advantage of the proposed FDL in accuracy and efficiency.

**Keywords:** Facial disguise learning, robust face recognition, regularized robust coding.

## 1 Introduction

As one of the most visible and challenging problems in computer vision and pattern recognition, face recognition (FR) has been extensively studied in the past two decades [5], and many representative methods, such as Eigenfaces [6], Fisherfaces [6], LBP [7], have been proposed. In order to deal with facial occlusion, Eigenimages [8-9] and probabilistic local approaches [10] were proposed for FR with occlusion. Although much progress have been made, robust FR to occlusion/disguise is still a challenging issue due to the variations of occlusion such as different categories of disguises, and the unknown intensity of occluded pixels.

Recently, sparse representation based classifier (SRC) [1] was proposed for robust face recognition, producing very promising performance in FR with occlusion. By coding a query image  $\mathbf{y}$  as a sparse linear combination of all the training samples via Eq. (1), SRC classifies  $\mathbf{y}$  by searching for the class that produces the minimal reconstruction error.

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \quad (1)$$

where  $\|\cdot\|_1$  is the sparse  $l_1$ -norm and each column vector in  $\mathbf{X}$  is a training sample. In order to make SRC robust to facial occlusion, an identity matrix  $\mathbf{I}$  was introduced as a dictionary to code the outlier pixels (e.g., occluded pixels):

$$\min_{\alpha e} \|y - X\alpha - Ie\|_2^2 + \lambda \|\alpha\|_1 + \lambda \|e\|_1 \quad (2)$$

By solving Eq. (2), SRC shows good robustness to face occlusions such as block occlusion and disguise. It is easy to see in Eq. (2) that the representation residual, i.e.,  $e$ , is regularized by  $l_1$ -norm, which may not be optimal if the representation residuals do not follow a Laplacian distribution. Following SRC, He *et al.* [11] proposed a correntropy-based sparse representation (CESR) for robust face recognition, which introduced a Gaussian kernel-based fidelity term to regularize the coding residuals; and Gabor feature was also introduced in the framework of SRC to enhance its discrimination [12]. In order to deal with more general facial occlusion, Yang *et al.* [4] proposed a regularized robust coding (RRC) model by designing a robust representation term, which has shown the state-of-art performance in robust face recognition and attracted much attention in the field.

Although RRC [3], CESR [11] and Gabor-SRC [12] have achieved leading performance in robust face recognition, all of them ignore to use the prior knowledge of facial disguise to further improve the performance. In practical face recognition, the commonest facial occlusion is the disguise, of which the pattern could be collected offline and used to advance the performance of FR. In this paper, we proposed a facial disguise learning (FDL) model, in which the unknown occlusion pattern in the query image is learned based on a disguise mask dictionary. Independent FDL and joint FDL algorithms were proposed with an efficient iterative solver for learning the query facial disguise. With the learned facial disguise, weighted sparse representation can be used for robust FR. We evaluate the effectiveness of FDL on AR [13] and a joint face database. The experiments on FR with disguises clearly show the advantage of FDL in accuracy and effectiveness of robust face recognition.

The rest of this paper is organized as follows. Section 2 briefly reviews the related regularized robust coding model. Section 3 presents the proposed facial disguise learning for face recognition. Section 4 conducts the experiments, and Section 5 concludes the paper.

## 2 Brief Review of Related Work

In order increase the robustness of SRC to various outliers, Yang et al [4] proposed a regularized robust coding (RRC) model, which was efficiently solved by using an iterative reweighted regularized coding algorithm. In each iteration RRC changes to

$$\min_{\alpha} \left\| \text{diag}(\mathbf{w})^{1/2} (y - X\alpha) \right\|_2^2 + \lambda \|\alpha\|_{l_p} \quad (3)$$

where  $l_p$ -norm on  $\alpha$  could be  $l_1$ -norm or  $l_2$ -norm ([2] indicated that the  $l_2$ -norm regularized coding could achieve similar accuracy to  $l_1$ -norm but with a faster speed), and  $\text{diag}(\mathbf{w})$  is a diagonal matrix with the weight vector  $\mathbf{w}$  as its diagonal vector. Here the element of  $\mathbf{w}$  is computed as

$$w_j = 1 / (1 + \exp(\mu e_j^2 - \mu \delta)) \quad (4)$$

where  $e_j = y_j - \mathbf{r}_j \boldsymbol{\alpha}$ ,  $\mathbf{r}_j$  is the  $j$ -th row vector of  $\mathbf{X}$ , and  $\mu$  and  $\delta$  are two automatically updated scalar parameters in the weight function [4]. Here  $w_j$  indicates the importance of the  $j$ -th element of  $\mathbf{y}$  to the coding of  $\mathbf{y}$ . We can observe that the outlier pixels will have small weights to reduce their effects on the coding  $\mathbf{y}$  on  $\mathbf{X}$  since they have big residuals.

When the final coding vector  $\boldsymbol{\alpha}$  is achieved, RRC conducts the classification via

$$\text{identity}(\mathbf{y}) = \arg \min_i \left\| \text{diag}(\mathbf{w})^{1/2} (\mathbf{y} - \mathbf{X}_i \boldsymbol{\alpha}_i) \right\|_2^2 \quad (5)$$

where  $\mathbf{X}_i$  the training samples of class  $i$ ,  $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1; \boldsymbol{\alpha}_2; \dots; \boldsymbol{\alpha}_c]$ , and  $\boldsymbol{\alpha}_i$  is the coefficient vector associated with  $i$ -th disguise pattern.

RRC could be solved by alternatively updating the weight vector  $\mathbf{w}$  and the coding vector  $\boldsymbol{\alpha}$ . The whole algorithm of RRC is briefly summarized in Table 1.

**Table 1.** Algorithm of RRC

<b>Solving algorithm of RRC</b>
1. Initialize $\boldsymbol{\alpha}$
2. Compute residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\alpha}$
3. Estimate weights $\mathbf{w}$ as via Eq.(4)
4. Solve $\boldsymbol{\alpha}$ via the weighted regularized robust coding , i.e., Eq.(3)
5. Output $\boldsymbol{\alpha}$ until the condition of convergence is met, or the maximal number of iterations is reached.

### 3 Facial Disguise Learning

In this section, we proposed a facial disguise learning (FDL) algorithm for robust face recognition. Fig. 1 shows some examples of facial disguise. As shown in the figure, although the disguise dramatically changes the appearance of face images (e.g., sunglasses with different color, scarves with different textures, hat with different styles, etc.), the occluded patterns are relatively stable (e.g., hat is on the head, scarf is below the nose, and sunglasses only cover the parts of eyes). Inspired by RRC, which use a weight vector (i.e.,  $\mathbf{w}$ ), to indicate the contribution of every pixel to the face image representation, we learn the query facial disguise  $\mathbf{w}$  from a collected facial disguise mask dictionary instead of the disguise itself with unknown intensities.



**Fig. 1.** Face images with disguises and the disguise masks

Given a set of collected facial disguise masks  $\mathbf{D}=[\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n]$ , where  $\mathbf{D}_j$  is  $j$ -th category of disguise masks. Given query  $\mathbf{y}$ , we want to estimate its weight vector (i.e.,  $\mathbf{w}$ ) to do robust regularized coding. In this paper, we proposed two facial disguise learning methods: independent FDL and joint FDL.

### 3.1 Independent Facial Disguise Learning (I-FDL)

I-FDL is used as a part of weight updating in the framework of RRC. We re-estimate the weight vector  $\mathbf{w}$  after the step 3 of the algorithm of RRC listed in Table 1.

Let  $\mathbf{w}$  be the weight vector estimated by step 3 in each iteration of RRC. To exploit the relation between  $\mathbf{w}$  and the prior knowledge of facial disguises, we firstly encode  $\mathbf{w}$  on the collected disguise mask dictionary  $\mathbf{D}$  as

$$\min_{\boldsymbol{\beta}} \|\mathbf{w} - \mathbf{D}\boldsymbol{\beta}\|_p + \kappa \|\boldsymbol{\beta}\|_p \quad (6)$$

where  $\kappa$  is a scalar parameter, and  $l_p$ -norm is used to regularize the reconstruction error and coding coefficients. After solving Eq.(6), the new estimated weight vector  $\mathbf{w}_n$  could be computed as

$$\mathbf{w}_n = \rho \mathbf{w} + (1-\rho) \mathbf{D}\boldsymbol{\beta} \quad (7)$$

where  $\rho$  is a scalar variable to make a balance between the original weight vector and new estimated weight vector. Then with  $\mathbf{w}_n$  in each iteration, I-FDL outputs the identity of  $\mathbf{y}$  by solving the algorithm of RRC. In our paper, we set  $\rho=0.5$ . The updating of  $\mathbf{w}$  using the prior facial disguise of  $\mathbf{D}\boldsymbol{\beta}$  could avoid the estimated  $\mathbf{w}_n$  to be arbitrary to some extent. For simplify in this paper the  $l_p$ -norm in Eq. (6) is  $l_2$ -norm. We can observe that Eq.(6) could be solved very fast.

### 3.2 Joint Facial Disguise Learning (J-FDL)

Different from I-FDL, we want to design a new model of robust face recognition by jointly learning the facial disguise and encoding the query face image. The proposed joint facial disguise learning (J-FDL) model is

$$\min_{\boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\beta}} \left\| \text{diag}(\mathbf{w})(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \right\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_p + \gamma (\|\mathbf{w} - \mathbf{D}\boldsymbol{\beta}\|_p + \kappa \|\boldsymbol{\beta}\|_p) \text{ s.t. } \sum_j \beta_j = 1 \quad (8)$$

where  $\gamma$  is a scalar to balance the facial disguise learning and query face image encoding, and the constraint  $\sum_j \beta_j = 1$  aims to avoid the trivial solution of  $\mathbf{w}$  (i.e.,  $\mathbf{w}=\mathbf{0}$ ). The regularizations on  $\mathbf{w}-\mathbf{D}\boldsymbol{\beta}$  and  $\boldsymbol{\beta}$  could be sparse  $l_1$ -norm or dense  $l_2$ -norm. For simplicity, in our paper, we set  $l_p$ -norm of  $\mathbf{w}-\mathbf{D}\boldsymbol{\beta}$  and  $\boldsymbol{\beta}$  regularized by  $l_2$ -norm. Therefore the J-FDL model could be written as

$$\min_{\boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\beta}} \left\| \text{diag}(\mathbf{w})(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \right\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_p + \gamma (\|\mathbf{w} - \mathbf{D}\boldsymbol{\beta}\|_2^2 + \kappa \|\boldsymbol{\beta}\|_2^2) \text{ s.t. } \sum_j \beta_j = 1 \quad (9)$$

We solve J-FDL by alternatively optimizing two convex sub-problem: robust coding of query sample (i.e., solving  $\alpha$ ) and facial disguise learning (i.e., solving  $w$  and  $\beta$ ). When  $w$  and  $\beta$  are fixed, the J-FDL becomes

$$\min_{\alpha} \|\text{diag}(w)(y - X\alpha)\|_2^2 + \lambda \|\alpha\|_p \quad (10)$$

which is a standard sparse coding or collaborative representation problem [2]. It can be solved by using the similar solver in RRC (e.g., step 4 listed in Table 1).

Denote  $e = y - X\alpha$ . When the coding coefficient  $\alpha$  is fixed, the proposed J-FDL changes to

$$\min_{w, \beta} \|\text{diag}(w)e\|_2^2 + \gamma \|w - D\beta\|_2^2 + \gamma\kappa \|\beta\|_2^2 \text{ s.t. } \sum_j \beta_j = 1 \quad (11)$$

from which we can easily derive that

$$w = \text{diag}(u)D\beta \quad (12)$$

where  $u_i = \gamma(\gamma e_i^2)$ . With Eq. (12), Eq. (11) changes to

$$\min_{\beta} \|I_{D1}\beta\|_2^2 + \gamma \|I_{D2}\beta\|_2^2 + \gamma\kappa \|\beta\|_2^2 \text{ s.t. } \sum_j \beta_j = 1 \quad (13)$$

where  $I_{D1} = \text{diag}(e)\text{diag}(u)$  and  $I_{D2} = \text{diag}(u)D - D$ .

Using Langrange multiplier, Eq.(13) is equivalent to

$$\min_{\beta, \tau} \|I_{D1}\beta\|_2^2 + \gamma \|I_{D2}\beta\|_2^2 + \gamma\kappa \|\beta\|_2^2 + \tau \left( \sum_j \beta_j - 1 \right) \quad (14)$$

Differentiating the objective function with respect to  $\beta$ , and let it be 0, we have

$$\beta = -\tau \left( 2I_{D1}^T I_{D1} + 2\gamma I_{D2}^T I_{D2} + 2\gamma\kappa I \right)^{-1} \mathbf{1} \quad (15)$$

where  $I$  is an identity matrix, and  $\mathbf{1}$  is a column vector with all elements as 1. So the solution of Eq. (11) under the constraint  $\sum_j \beta_j = 1$  is

$$\beta = \beta / \sum_j \beta_j \quad (16)$$

After solving  $w$  and  $\beta$ , the coding vector  $\alpha$  could be updated. Through several iteration, we could get the final weight vector  $w$  and coding vector  $\alpha$ , and then conduct face recognition via

$$\text{identity}(y) = \arg \min_i \|\text{diag}(w)(y - X_i \alpha_i)\|_2^2 \quad (17)$$

### 3.3 Disguise Recognition

Apart from FR, disguise recognition could be conducted by

$$\text{disguise} = \arg \min_i \|w - D_i \beta_i\|_2^2 \quad (18)$$

where  $\beta = [\beta_1; \beta_2; \dots; \beta_n]$  and  $\beta_i$  is the coefficient vector associated with i-th disguise pattern.

## 4 Experiments

We perform experiments on AR database [13] and a joint database to demonstrate the performance of JDL. In Section 4.1, we test FDL on AR database; in Section 4.2, we test FDL on a joint database; and finally Section 4.3 presents the comparison of running time. Here the joint database was constructed by using AR database (100 persons, 2599 images) [13] and a subset of CAS-Peal (101 persons and 843 images) [14]. We randomly select 60 images with sunglasses or scarf of 15 subjects from AR and 61 images with hat from CAS-Peal to construct the disguise mask dictionary. Here we manually segment the facial disguises to collect the disguise mask dictionary; and the values of occluded pixels are set as 0, with the values of all the other pixels as 1.

All the face images are cropped to the size of  $42 \times 30$  and aligned by using the locations of eyes. We normalize the query image and training image to have unit  $l_2$ -norm energy. In all experiments  $\kappa$  of FDL is empirically set as 0.5. For J-FDL there is another parameter,  $\gamma$ , which was set as 50 in face recognition and 1000 in disguise recognition. The competing methods include the latest approaches, such as SRC [1], Gabor-SRC [12], CESR [11], RRC\_L1 [3] and RRC\_L2[3]. Similar to RRC, FDL\_L1 and FDL\_L2 represents FDL using  $l_1$ -norm and  $l_2$ -norm on  $\alpha$ , respectively.

### 4.1 Face Recognition on AR Database

A subset from the AR database [13] is used in this experiment. This subset consists of 2,599 images from 100 subjects (26 samples per class except for a corrupted image w-027-14.bmp), 50 males and 50 females. As [1], 799 images (about 8 samples per subject) of non-occluded frontal views with various facial expressions in Sessions 1 and 2 were used for training, while two separate subsets (with sunglasses and scarf) of 200 images (1 sample per subject per Session, with neutral expression) were used for testing. The FR results by the competing methods are listed in Table 2. We can see that FDL methods achieve higher recognition rates than the second best methods, RRC. For instance, J-FDL\_L1 outperforms RRC\_L1, RRC\_L2, GSRC, SRC and CESR by 1.5%, 2.5%, 20%, 39.5% and 57% on FR with scarf, respectively. We can also see that J-JDL is better than I-JDL, which shows the advantage of joint facial disguise learning and face representation. The proposed FDL methods also significantly outperform other state-of-the-art methods, including [3] with 84% on sunglasses and 93% on scarf, and [16] with 93% on sunglasses and 95.5% on scarf.

### 4.2 Face Recognition on a Joint Database

In the test, we conduct FR with more complex disguises (e.g., sunglasses, scarf and hat) with variations of illumination and longer data acquisition interval. Apart from the subjects for constructing the disguise mask dictionary, 340 images of the remaining 85 subjects (4 natural and non-occluded images with different illuminations in

Session 1) in AR database and 263 images of the remaining 80 subjects (the non-occluded images) in CAS-Peal are used as the training sets. And 510 face images with sunglass and lighting variations, 510 face images with scarf and lighting variations, and 240 face images with hat and lighting variations are used as the testing dataset. Some samples are shown in Fig. 2.

**Table 2.** Recognition rates of the AR database with facial disguise

Algorithms	Sunglasses	Scarves
SRC	87.0%	59.5%
GSRC	93%	79%
CESR	99%	42.0%
RRC_L <sub>2</sub>	99.5%	96.5%
<b>I-FDL_L2</b>	<b>100%</b>	96.5%
<b>J-FDL_L2</b>	<b>100%</b>	98.0%
RRC_L <sub>1</sub>	<b>100%</b>	97.5%
<b>I-FDL_L1</b>	<b>100%</b>	97.5%
<b>J-FDL_L1</b>	<b>100%</b>	<b>99.0%</b>

We first show the results of disguise recognition in Table 3 by comparing the proposed FDL and RRC. From Table 3 we can observe that FDL and RRC have similar disguise recognition accuracy in the cases of sunglasses and hat. However, in the case of scarf, FDL is significantly better than RRC (e.g., about 20% improvement for I-FDL\_L2 and 27% improvement for J-FDL\_L2). It seems that scarf is more challenging due to its big size and RRC can not well recover it.



**Fig. 2.** The training and testing samples in the joint database

Table 4 lists the results of face recognition on the joint database by competing methods. Clearly, the FDL methods achieve much better results than SRC, GSRC, CESR and RRC in most cases. RRC achieves the second best performance. J-FDL\_L2 and I-FDL\_L2 outperforms RRC\_L2 by 10% and 4.6% in average, respectively; J-FDL\_L1 and I-FDL\_L1 outperforms RRC\_L1 by 5% and 2.5% in average, respectively. It is also interesting that FDL works better than RRC in the challenging face recognition with hat disguise although RRC recognizes hat slightly better than FDL. It may be because the estimated weight vector of FDL is more suitable for face recognition than that of RRC.

**Table 3.** Disguise recognition rates on the joint database with disguise

Algorithms	Sunglass	Scarf	Hat
RRC_L <sub>2</sub>	99.4%	70.5%	<b>99.6%</b>
<b>I-FDL_L2</b>	<b>100%</b>	89.2%	98.3%
<b>J-FDL_L2</b>	<b>100%</b>	<b>97.5%</b>	94.6%
RRC_L <sub>1</sub>	99.8%	69.6%	<b>99.2%</b>
<b>I-FDL_L1</b>	<b>100%</b>	89.6%	97.9%
<b>J-FDL_L1</b>	<b>100%</b>	<b>97.8%</b>	95.4%

**Table 4.** Recognition rates on the joint database with three facial disguises

Method	Sunglass	Scarf	Hat
SRC	73.9%	24.9%	26.3%
GSRC	52.4%	66.1%	34.2%
CESR	80.2%	11.0%	26.7%
RRC_L <sub>2</sub>	83.5%	75.3%	60.4%
<b>I-FDL_L2</b>	89.4%	81.2%	62.5%
<b>J-FDL_L2</b>	<b>91.8%</b>	<b>82.0%</b>	<b>75.8%</b>
RRC_L <sub>1</sub>	90.2%	77.3%	67.1%
<b>I-FDL_L1</b>	<b>93.3%</b>	<b>85.5%</b>	63.3%
<b>J-FDL_L1</b>	92.7%	83.7%	<b>73.8%</b>

### 4.3 Running Time Comparison

Apart from recognition rate, computational expense is also an important issue for practical FR systems. In this section, the running time of the baseline method, SRC, and some competing methods which show not bad performance in all cases, including GSRC, RRC\_L2, RRC\_L1, and FDL, is evaluated using the FR experiments on the joint face database. The programming environment is Matlab version R2013a. The desktop used is equipped with a 3.5 GHz CPU and 16G RAM. All the methods are implemented using the codes provided by the authors. For SRC, we use a fast  $l_1$ -minimization solver, ALM [15], to implement the sparse coding step.

The recognition rates have been reported in Table 4. Table 5 lists the average computational expense of different methods. We can observe that both FDL\_L2 and RRC\_L2 have the least running time, followed by GSRC and SRC. Although the proposed FDL has similar computation time to RRC, FDL could achieve visibly better performance than RRC. Especially, FDL\_L2 has very small running time, but much better accuracy than previous methods.

**Table 5.** Average runnning time on the joint database with three facial disguises

Method	Sunglass	Scarf	Hat
SRC (ALM)	0.610	0.579	0.574
GSRC	0.269	0.265	0.277
RRC_L <sub>2</sub>	0.177	0.153	0.171
<b>I-FDL_L2</b>	0.142	0.167	0.174
<b>J-FDL_L2</b>	0.186	0.172	0.192
RRC_L <sub>1</sub>	1.58	1.34	1.59
<b>I-FDL_L1</b>	0.963	1.17	1.09
<b>J-FDL_L1</b>	1.07	1.23	1.11

## 5 Conclusion

This paper presented a novel facial disguise learning (FDL) model for robust face recognition and an associated effective iterative algorithm. One important advantage of FDL is that it could automatically learn the disguise pattern of the query image by using a disguise mask dictionary. We proposed two ways to fully exploit the prior knowledge of facial disguise. The independent FDL only re-estimate the weight of pixels in the scheme of RRC; while the joint FDL is a new model, which jointly learns the disguise pattern and the representation of the query image. Apart from robust FR, we also showed that FDL could be used in disguise recognition. The proposed FDL methods were extensively evaluated on FR with various facial disguises. The experimental results clearly demonstrated that FDL outperforms significantly previous state-of-the-art methods, such as SRC, CESR, GSRC and RRC. In particular, FDL with  $\ell_2$ -norm regularization could achieve very high recognition rates but with low computational cost, which makes it a very good candidate scheme for practical robust FR systems.

**Acknowledgement.** This work is partially supported by the National Natural Science Foundation of China under Grants no. 61402289 and 61272050, and Shenzhen Scientific Research and Development Funding Program under Grants JCYJ20140506231950374 and JCYJ20130329115750231.

## References

1. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)
2. Zhang, L., Yang, M., Feng, X.C.: Sparse representation or collaborative representation: which helps face recognition? In: Proc. ICCV (2011)
3. Fidler, S., Skocaj, D., Leonardis, A.: Combining Reconstructive and Discriminative Subspace Methods for Robust Classification and Regression by Subsampling. *IEEE Trans. Pattern Analysis and Machine Intelligence* 28(3), 337–350 (2006)
4. Yang, M., Zhang, L., Yang, J., Zhang, D.: Regularized robust coding for face recognition. *IEEE Trans. Image Processing* 22(5), 1753–1766 (2013)
5. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. *ACM Computing Survey* 35(4), 399–458 (2003)
6. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(7), 711–720 (1997)
7. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 28(12), 2037–2041 (2006)
8. Leonardis, A., Bischof, H.: Robust recognition using eigenimages. *Computer Vision and Image Understanding* 78(1), 99–118 (2000)
9. Chen, S., Shan, T., Lovell, B.C.: Robust face recognition in rotated eigenspaces. In: Proc. Int'l Conf. Image and Vision Computing, New Zealand (2007)

10. Martinez, A.M.: Recognizing Imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(6), 748–763 (2002)
11. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 33(8), 1561–1576 (2011)
12. Yang, M., Zhang, L.: Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*, Part VI. LNCS, vol. 6316, pp. 448–461. Springer, Heidelberg (2010)
13. Martinez, A., Benavente, R.: The AR face database. CVC Tech. Report No. 24 (1998)
14. Gao, W., Cao, B., Shan, S.G., Chen, X.L., Zhou, D.L., Zhang, X.H., Zhao, D.B.: The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. *IEEE Trans. on System Man, and Cybernetics (Part A)* 38(1), 149–161 (2008)
15. Yang, A.Y., Ganesh, A., Zhou, Z.H., Sastry, S.S., Ma, Y.: A review of fast l1-minimization algorithms for robust face recognition. *arXiv:1007.3753v2* (2010)
16. Jia, H., Martinez, A.: Support vector machines in face recognition with occlusions. In: *Proc. CVPR* (2009)

# A Static Hand Gesture Recognition Algorithm Based on Krawtchouk Moments

Shuping Liu<sup>1</sup>, Yu Liu<sup>1</sup>, Jun Yu<sup>1,2</sup>, and Zengfu Wang<sup>1,2,3</sup>

<sup>1</sup> Department of Automation, University of Science and Technology of China,  
Hefei 230026, China

<sup>2</sup> National Engineering Laboratory for Speech and Language Information Processing,  
University of Science and Technology of China, Hefei 230026, China

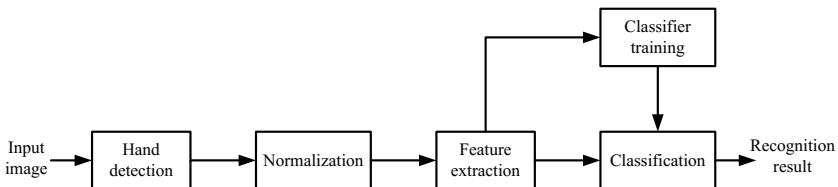
<sup>3</sup> Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China  
[{fengya.liuyu1@mail.ustc.edu.cn}](mailto:{fengya.liuyu1@mail.ustc.edu.cn}), [{harryjun,zfwang}@ustc.edu.cn](mailto:{harryjun,zfwang}@ustc.edu.cn)

**Abstract.** Owing to convenience and naturalness, hand gesture recognition has been widely used in various human-computer interaction (HCI) systems. In this paper, we address the problem from the perspective of system, and present a static hand gesture recognition algorithm based on Krawtchouk moments. The effect of the order and number of Krawtchouk moments on the recognition performance is studied in detail. In the experiments, 15 popular gesture signs are used to verify the effectiveness of the presented hand gesture recognition system. Experimental results demonstrate that lower order Krawtchouk moments are more suitable for classification. Furthermore, the number of Krawtchouk moments also has a significant impact on the recognition accuracy.

**Keywords:** hand gesture recognition, hand detection and normalization, Krawtchouk moments, minimum distance classifier.

## 1 Introduction

In recent years, hand gesture has emerged as an important communication modality in human-computer interaction (HCI). Compared with the traditional interfaces such as keyboard and mouse, hand gesture owns clear advantages in HCI for its convenience and naturalness. Hand gesture interfaces have been successfully used in many applications such as robotics, sign language communication, video annotations, assistive systems, virtual reality, etc. In general, hand gesture recognition can be classified into two groups: static hand gesture (hand posture) recognition and dynamic hand gesture recognition. Static hand gesture recognition usually focuses on the hand shape in a still image, while the movement of hand tends to be mainly taken into consideration in dynamic hand gesture recognition. Both two types of hand gesture recognition have practical meanings in HCI since they have different emphases and applications. In this work, we pay our attention to 2D image based static hand gesture recognition, which is more difficult to some extent since the information can only be extracted from a still image.



**Fig. 1.** A typical flow chart of a static hand gesture recognition system

Fig. 1 shows a typical flow chart of a static hand gesture recognition system. The target of hand detection is to localize the hand in the input image. Then, some normalization operations (such as rotation, translation and scale normalization) should be performed on the detected hand region. The next step is feature extraction, which aims to make a quantitative representation for the normalized hand. Finally, the input hand gesture is classified into a certain category using a trained classifier, which is obtained with a pre-designed classifier model and some training examples.

Among all the modules in Fig. 1, the feature extraction stage, which has a great impact on the final recognition accuracy, plays a very crucial role in the hand gesture recognition system. Feature extraction may be based on either gray image or binary image. For gray image based feature extraction, the features can be extracted from the pixels' values by some dimension reduction techniques such as principal component analysis (PCA) [1, 2], local invariant features such as scale invariance feature transform (SIFT) [3, 4], and elastic graph matching technique [5, 6]. For binary image based feature extraction, the hand region needs to be accurately segmented from the input image. In particular, a hand region can be quantitatively represented by its contour or the whole region, namely, contour based feature extraction [7] and region based feature extraction [8]. Popular contour based feature representation and description techniques [9] include Fourier descriptors, chain code representations, curvature scale space, etc. However, these contour based approaches are usually sensitive to noise and contour distortions. The region based feature representations are calculated by considering the entire pixels belonging to the hand region. Convex hulls, shape matrices and moment invariants are commonly used region based approaches [9]. The analysis in [9] shows that the moments based feature representations own some advantages such as compact representation, robustness to noise, invariance properties, and low computational as well as storage costs. Conventionally, the geometric moments such as Hu's moment invariants [10] are widely employed in shape analysis and object recognition. However, these moments are non-orthogonal [11], so the information contained in the moments is redundant and reconstructing the image from the moments is not possible. To overcome this problem, many orthogonal moments have been introduced into image analysis and achieved great success, such as the Zernike moments [11], Tchebichef moments [12] and Krawtchouk moments [13]. Recently, Priyal and Bora [8] applied the Krawtchouk moments to hand gesture recognition.

Their experimental results show the clear advantages of Krawtchouk moments over other moments like the Zernike moments used for gesture recognition. However, they paid most of their attention to the hand detection and normalization stages (also very important) in [8], while the effect of Krawtchouk moments on the recognition performance was not deeply investigated. Furthermore, there are too many empirical or experimental parameters in their detection and normalization methods, which also limit the usefulness to a large extent.

In this paper, based on Krawtchouk moments, we present a complete static hand gesture recognition system, which includes hand detection, normalization, feature extraction and gesture recognition. In the experiments, 15 popular gesture signs are employed to verify the effectiveness of the presented hand gesture recognition system. Furthermore, the impact of some parameters such as the order and number of Krawtchouk moments on the recognition performance is studied in detail.

## 2 Krawtchouk Moments

Krawtchouk moments are a set of discrete orthogonal moments which are derived from the Krawtchouk polynomials [14]. In this section, we make a brief introduction to Krawtchouk polynomials and Krawtchouk moments.

### 2.1 Krawtchouk Polynomials

The  $n$ -th order classical Krawtchouk polynomial at a discrete point  $x$  with  $(0 < p < 1)$  is defined as

$$K_n(x; p, N) = {}_2F_1(-n, -x; -N; \frac{1}{p}), \quad (1)$$

where  $x = 0, 1, \dots, N$ ,  $N > 0$ .  ${}_2F_1$  is a hypergeometric function defined as

$${}_2F_1(a, b; c; z) = \sum_{v=0}^{\infty} \frac{(a)_v (b)_v}{(c)_v} \frac{z^v}{v!}, \quad (2)$$

where  $(a)_v = a(a+1)\cdots(a+v-1)$  is the Pochhammer symbol. Examples of Krawtchouk polynomial up to second order are

$$K_0(x; p, N) = 1, \quad K_1(x; p, N) = 1 - \frac{x}{Np}, \quad K_2(x; p, N) = 1 - \frac{x}{Np}(2 - \frac{x-1}{(N-1)p}).$$

The set of  $(N + 1)$  Krawtchouk polynomials  $\{K_n(x; p, N)\}$  forms a complete orthogonal basis with a binomial weight function

$$w(x; p, N) = \binom{N}{x} p^x (1-p)^{N-x}. \quad (3)$$

The orthogonality property is given by

$$\sum_{x=0}^N w(x; p, N) K_n(x; p, N) K_m(x; p, N) = \rho(n; p, N) \delta(n - m), \quad (4)$$

where  $n, m = 1, 2, \dots, N$ ,  $\delta(\cdot)$  is the Kronecker delta function, and

$$\rho(n; p, N) = (-1)^n \left( \frac{1-p}{p} \right)^n \frac{n!}{(-N)_n}. \quad (5)$$

## 2.2 Weighted Krawtchouk Polynomials

Yap et al. [13] experimentally found that the range of values of the polynomials expands rapidly with a slight increase of the order, which cannot ensure the stability of the Krawtchouk polynomials. Therefore, a set of weighted Krawtchouk polynomials  $\{\bar{K}_n(x; p, N)\}$  are introduced as the following definition

$$\bar{K}_n(x; p, N) = K_n(x; p, N) \sqrt{\frac{w(x; p, N)}{\rho(n; p, N)}}, \quad (6)$$

so the orthogonality condition in (4) becomes

$$\sum_{x=0}^N \bar{K}_n(x; p, N) \bar{K}_m(x; p, N) = \delta(n - m). \quad (7)$$

The parameter  $p$  can be viewed as a translation factor. If  $p = 0.5 + \Delta p$ , the weighted Krawtchouk polynomials are shifted by about  $N\Delta p$ . The direction of shifting relies on the sign of  $\Delta p$ , with the polynomials shifting along  $+x$  direction when  $\Delta p$  is positive and vice versa.

In practice, the Krawtchouk polynomials are calculated recursively with the following relation

$$p(n-N-1)K_n(x) = (x+1-2p+2pn-n-Np)K_{n-1}(x) - (p-1)(n-1)K_{n-2}(x). \quad (8)$$

Then, with (7), we can obtain the relation of weighted Krawtchouk polynomials

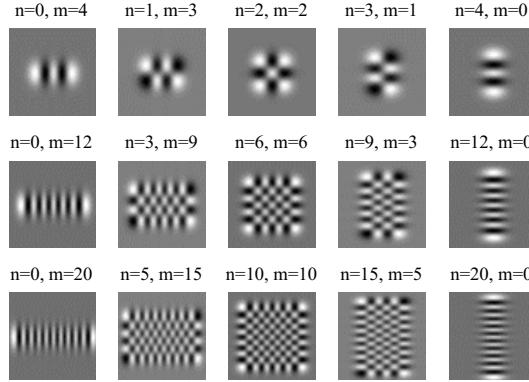
$$p(n-N-1)\bar{K}_n(x) = A(x+1-2p+2pn-n-Np)\bar{K}_{n-1}(x) - B(p-1)(n-1)\bar{K}_{n-2}(x), \quad (9)$$

where

$$A = \sqrt{\frac{p(N-n+1)}{(1-p)n}}, \quad B = \sqrt{\frac{p^2(N-n+1)(N-n+2)}{(1-p)^2(n-1)n}},$$

with

$$\bar{K}_0(x; p, N) = \sqrt{w(x; p, N)}, \quad \bar{K}_1(x; p, N) = (1 - \frac{x}{pN}) \sqrt{\frac{w(x; p, N)}{\rho(1; p, N)}}.$$



**Fig. 2.** Some examples of 2D Krawtchouk basis for  $N = M = 64, p_1 = p_2 = 0.5$

### 2.3 Krawtchouk Moments

Based on the weighted Krawtchouk polynomials, the  $(n+m)$ th order Krawtchouk moments of image  $f(x, y)$  of size  $N \times M$  is defined as

$$Q_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{K}_n(x; p_1, N) \bar{K}_m(y; p_2, M) f(x, y). \quad (10)$$

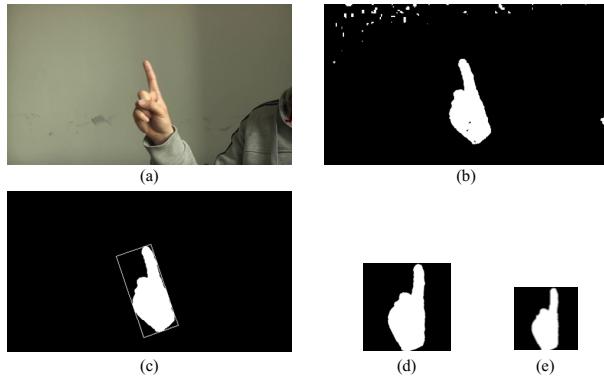
Fig. 2 shows some examples of 2D Krawtchouk basis for  $N = M = 64, p_1 = p_2 = 0.5$  with different  $n$  and  $m$ . It can be seen that when the order becomes higher, the polynomials have wider supports. Thus, the lower order Krawtchouk moments capture local features while the higher order ones mainly represent the global characteristics.

## 3 Hand Gesture Recognition System

### 3.1 Hand Detection

In this work, we use the skin color model proposed in [15] by Teng et al. to detect the hand in the input image. As with the assumption in [8], we restrict the background so that the hand is the largest region with respect to the skin color (a method to eliminate the disturbance of face by face subtraction is presented in [4]). Moreover, the forearm is assumed to be full dressed since this work mainly concentrates on the feature extraction stage (a detailed forearm detection method is reported in [8]). In [15], the skin color regions are determined by combining the information obtained from YCbCr and YIQ color spaces. The hue value  $\theta$  is calculated using the Cb-Cr chromatic components by

$$\theta = \tan^{-1} \left( \frac{Cr}{Cb} \right). \quad (11)$$



**Fig. 3.** (a) Input hand gesture image. (b) The skin color detection result with (13). (c) The segmented result. (d) The detection result after rotation and translation normalization. (e) The final normalized detection result.

The in-phase color component  $I$  is obtained by a linear combination of RGB components

$$I = 0.596R - 0.274G - 0.322B. \quad (12)$$

According to [15], for Asian and European skin tones, the pixels are viewed as skin color pixels if they simultaneously satisfy

$$105^\circ \leq \theta \leq 150^\circ, \quad 30 \leq I \leq 100. \quad (13)$$

Fig. 3(b) shows the skin color detection result of the input hand gesture image shown in Fig. 3(a) using (13). We can see that the detection result not only contains some other objects which do not belong to the hand, but also misses some real hand pixels. Thus, a postprocessing technique is employed to just preserve the largest region as well as fill the small holes in the hand region. Furthermore, we also apply the morphological closing operation with a disk-shaped structuring element to obtain a smoother segmented result. Fig. 3(c) shows the segmented result of Fig. 3(a).

### 3.2 Normalization

As the Krawtchouk moments in (10) are not invariant to rotation, translation and scale, several appropriate normalization techniques are employed to improve the final recognition efficiency.

#### Rotation Normalization

We use the minimum enclosing rectangle to normalize the rotation. Minimum enclosing rectangle is defined as the smallest rectangle that can enclose the connected target region. As shown in Fig. 3(c), it generally reflects the orientation of the target region. Particularly, the orientation of the longer side with respect to the vertical axis is regarded as the rotation angle of the hand gesture.

### Translation Normalization

After rotation normalization, the normalization of the spatial position of the hand gesture is an easy task. We first find the center point of the rotated enclosing rectangle and employ it as the reference point. Then, we fix the center point and extend the rectangle region to a square with its side length equaling to the longer side length of the rectangle. Fig. 3(d) illustrates the detection result of Fig. 3(a) after translation normalization.

### Scale Normalization

At last, the spatial size of the hand gesture region is normalized to a fixed size such as  $64 \times 64$  through bilinear interpolation technique. Fig. 3(e) shows the final normalized detection result of Fig. 3(a).

### 3.3 Feature Extraction and Classification

For the normalized gestures, their Krawtchouk moments with different orders calculated with (10) are employed to form the feature vectors for training and classification. In this work, we use the minimum distance classifier to classify different hand gestures. Let  $v_p$  and  $v_q$  denote the feature vector of a test image and a training image, respectively. The Euclidean distance is used to measure the similarity between the two gestures. Thus, the most matched example in training set is obtained by

$$q^* = \arg \min_q (\|v_p - v_q\|_2^2), \quad q \in Q, \quad (14)$$

where  $q$  is the label of a gesture in the training set and  $Q$  is the set which contains all possible labels.

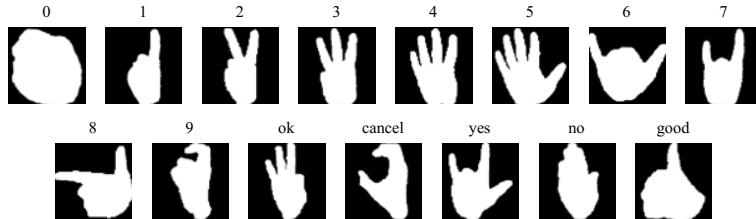
## 4 Experiments

### 4.1 Database

In our experiments, 15 popular gesture signs as shown in Fig. 4 are employed to verify the effectiveness of the presented recognition system. As is well known, the robustness to user variations is a very important factor in gesture recognition, so 12 users are invited to take part in the experiments. For each gesture sign of each user, we evenly collect 50 gesture images of size  $480 \times 270$  pixels from a short video, in which the rotation angle varies from  $-60^\circ$  to  $60^\circ$  with respect to the vertical axis. Among the 50 gesture images, 30 of them are used for training and the remaining 20 are used to test the classification performance. Therefore, for each gesture sign, there are 360 training examples and 240 test examples.

### 4.2 Experimental Results and Discussions

In this subsection, we mainly study the impact of the orders of Krawtchouk moments on the recognition accuracy. Considering that the centrosymmetric 2D

**Fig. 4.** Gesture signs in our experiments**Table 1.** The average recognition accuracies of 15 gesture signs with different groups of Krawtchouk moments

Dimension	4		8		12	
Orders	1,2,3,4	2,4,6,8	1,2,...,8	2,4,...,16	1,2,...,12	2,4,...,24
Accuracy	81.48%	75.89%	95.67%	86.67%	96.89%	87.67%

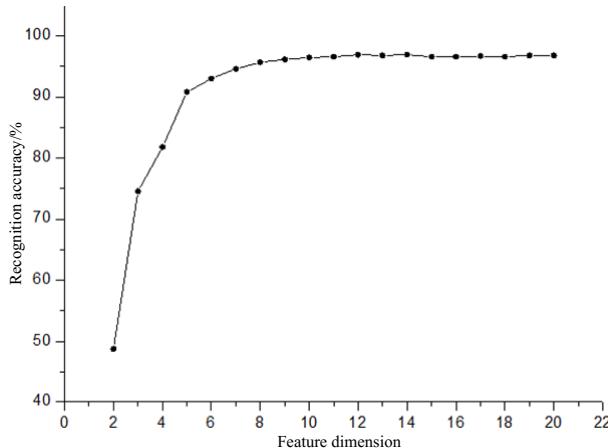
Krawtchouk moments are more suitable for feature representation of a hand region, we set  $p_1 = p_2 = 0.5$  and  $n = m$  for all the moments in this work.

We first fix the dimension of features (number of selected Krawtchouk moments), and compare the performance of Krawtchouk moments with different orders. Specifically, the feature dimensions are set to 4, 8 and 12, respectively. For each dimension of  $k$ , two groups of Krawtchouk moments are used to make classification. The orders of Krawtchouk moments are set to  $n = m = \{1, 2, \dots, k\}$  in the first group and  $n = m = \{2, 4, \dots, 2k\}$  in the second. Table 1 lists the average recognition accuracies of 15 gesture signs with different groups of Krawtchouk moments. It can be seen that for each dimension of 4, 8 or 12, the first group clearly outperform the second one in terms of recognition accuracy, which indicates that the low order Krawtchouk moments are more suitable for feature extraction.

Another realistic problem that arises from Table 1 is the selection of feature dimension. The recognition accuracy generally increases with higher dimensional feature. However, the computational and storage costs will increase when the dimension becomes higher. Thus, we conduct a set of experiments

**Table 2.** The average recognition accuracies of 15 gesture signs with different feature dimensions ranging from 2 to 20

Dimension	2	3	4	5	6	7	8
Accuracy	48.78%	74.56%	81.78%	90.78%	93.00%	94.56%	95.67%
Dimension	9	10	11	12	13	14	15
Accuracy	96.11%	96.44%	96.56%	96.89%	96.78%	96.89%	96.56%
Dimension	16	17	18	19	20		
Accuracy	96.56%	96.67%	96.56%	96.78%	96.78%		



**Fig. 5.** The relation between the feature dimension and recognition accuracy

to study the relation between the feature dimension and recognition accuracy. Based on the result in Table 1, we select the Krawtchouk moments with orders  $n = m = \{1, 2, \dots, k\}$  for each dimension of  $k$ . The dimension ranges from 2 to 20 with a step of 1. The average recognition accuracies of 15 gesture signs with different feature dimensions are listed in Table 2, from which we can see that the recognition accuracy tends to be stable when the dimension is larger than 10. The relation curve shown in Fig. 5 gives a more intuitive exhibition. Therefore, the feature dimension can be set to 10 to make a good balance between recognition accuracy and computational/storage cost. In our experiments, when the dimension is 10, it takes about 0.12 seconds to classify a gesture image on a computer with 3.0 GHz CPU and 4 GB RAM, so the presented gesture recognition system has the potential to be used in real-time applications.

## 5 Conclusions

In this paper, we present an efficient static hand gesture recognition system based on Krawtchouk moments. The hand region is first detected with a skin color model. Then, we introduce a novel technique for both rotation and translation normalization based on minimum enclosing rectangle. Finally, the Krawtchouk moments are used as the features for gesture classification. In the experiments, 15 popular gesture signs are employed to verify the effectiveness of the presented gesture recognition system. Experimental results demonstrate that lower order Krawtchouk moments are more suitable for feature representation, and the feature dimension (number of Krawtchouk moments) also has a significant impact on the recognition accuracy. We plan to further investigate the effect of some other parameters of the Krawtchouk moments on the recognition accuracy. Particularly, some Krawtchouk moments with different translation factors may be

combined into the feature vector to increase the distinction among different gesture signs, so that the performance of the hand gesture recognition system can be improved.

**Acknowledgements.** This work was supported by the National Natural Science Foundation of China (No. 61303150), the Anhui Province Initiative Funds on Intelligent Speech Technology and Industrialization (No. 13Z02008), and the Fundamental Research Funds for the Central Universities.

## References

1. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
2. Saxena, A., Jain, D.K., Singhal, A.: Sign language recognition using principal component analysis. In: International Conference on Communication Systems and Network Technologies, pp. 810–813 (2014)
3. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
4. Triesch, J., von der Malsburg, C.: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and Measurement 60, 3592–3607 (2011)
5. Dardas, N.H., Georganas, N.D.: Classification of hand postures against complex backgrounds using elastic graph matching. Image and Vision Computing 20, 937–943 (2002)
6. Li, Y.T., Wachs, J.P.: HEGM: A hierarchical elastic graph matching for hand gesture recognition. Pattern Recognition 47, 80–88 (2014)
7. Yao, Y., Fu, Y.: Contour model based hand-gesture recognition using Kinect sensor. IEEE Transactions on Circuits and Systems for Video Technology (2014)
8. Priyal, S.P., Bora, P.K.: A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments. Pattern Recognition 46, 2202–2219 (2013)
9. Zhang, D., Lu, G., Mitra, S.K.: Review of shape representation and description techniques. Pattern Recognition 37, 1–19 (2004)
10. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transaction on Information Theory IT-8, 179–187 (1962)
11. Teague, M.R.: Image analysis via the general theory of moments. Journal of Optic Society of America 70, 920–930 (1962)
12. Mukundan, R., Ong, S.H., Lee, P.A.: Image analysis by Tchebichef moments. IEEE Transactions on Image Processing 10, 1357–1364 (2001)
13. Yap, P.T., Paramesran, R., Ong, S.H.: Image analysis by Krawtchouk moments. IEEE Transactions on Image Processing 12, 1367–1376 (2003)
14. Krawtchouk, M.: On interpolation by means of orthogonal polynomials. Memoirs Agricultural Inst. Kyiv. 4, 21–28 (1929)
15. Teng, X., Wu, B., Yu, W., et al.: A hand gesture recognition based on local linear embedding. Journal of Visual Languages and Computing 16, 442–454 (2005)

# Face Recognition in the Wild by Mining Frequent Feature Itemset

Yuzhuo Wang, Hong Cheng, Yali Zheng, and Lu Yang

Center for Robotics and School of Automation Engineering, Chengdu 611731, China  
wangyuzhuo1989@gmail.com, {hcheng,zhengyl,yanglu}@uestc.edu.cn

**Abstract.** Face recognition has attracted a lot of attention in the last decades and achieved high recognition rate under controlled environment. More and more researchers now focus on face recognition in the wild, which is difficult because of the variance of pose, illumination, occlusion and so on. In this paper, we aim to solve this problem by combining image retrieval and feature weighting. By image retrieval method, we can find those face images in the gallery set which are the most similar to the probe face image. After getting similar face subset, feature weighting is then executed on this subset. This process includes two steps. In the first step, we learn a weight for each single feature in this subset by finding its nearest neighbor. In the second step, inspired by frequent item mining method we learn a weight for a group of features. In the testing process, by weighted nearest neighbor voting for both single and grouped features, we classify the probe image to the class which has the highest similarity score. We evaluate our method on AR and Pubfig83 face data sets. Experiment shows that our method has achieved state-of-the-art performance.

**Keywords:** face recognition, natural scenes, FIM, feature weighting.

## 1 Introduction

Face recognition has been well studied over the past decades because of its broad applications such as access control, public security, and human-computer interaction. We can divide face recognition task into two categories. One is to match an unknown face image to a specific class of a gallery of people, which is called face identification. The other one is face verification, which aims to decide whether two given images are from the same person or not. Obviously, the latter is much more easier than the former one. For face verification, there is a well-known database called LFW [14]. A lot of approaches have been developed and got significant performance on this database. So the method proposed in this paper mainly addresses the face identification problem in natural environment.

For the problem of face recognition in natural scenes, one of the most important challenge is that intro-class difference is often smaller than inter-class difference due to various illumination, occlusion, pose, expression and accessories. That means among all the face images in the gallery set, there are some

faces which are very similar to the probe image. And of course, there also are some faces which are different from the test one even though they come from the same class. Most face recognition methods calculate the similarity of test image with gallery set by image-to-image comparison, some of which are unnecessary since some face image are bearing different appearance with the probe image. To eliminate those similar ones is our main challenge. In this paper, we adopt image retrieval method proposed by Y. Wu *et al* [23] to get this similar face subset.

Most of face identification algorithms such as PCA [26], LDA [5] treat face recognition problem as a dimension reduction problem over the years, which consider the original face images in high dimensional space lying in a low dimensional manifold. It has been successfully applied to human identification, security access in the laboratorial scenes. In contrast with the controlled conditions, face images captured in the wild vary highly in hair fashion, making up, expression, illuminations, poses of individuals and times. Therefore, identifying a face in natural scenes is a more challenging task. And local feature is more suitable in this case.

Researchers have developed many local features to represent objects, such as raw patch, Scale-Invariant Feature Transform (SIFT) [18], Local Binary Pattern (LBP), Gabor wavelet, or combined features. The SIFT feature is known as one of the best descriptors so far, and has been widely and successfully used in object recognition and image retrieval. Yet, its ability has not been fully and deeply utilized in face recognition. In this paper, we utilize the location of features in each image following the SIFT descriptor extraction, called as spatial-SIFT, to capture the information of faces. We introduce a parameter  $\alpha$  to balance the influence of location term in SIFT feature matching process. This parametric spatial-SIFT feature can greatly eliminate the wrong matches.

Unlike the powerful classifiers SVM, Adaboost, nearest neighbor is a non-parameter algorithm without any training process, yet it is able to deal with a huge number of classes [6,9], and easy to be implemented in a parallel way for a large scale data. The rest of this paper is organized as follows. Section 2 reviews the related work in face identification and feature weighting. In section 3, we present our proposed image retrieval and feature weighting based face identification method in details. Experiment results with analysis and comparison with state-of-the-art methods are shown in Section 4. Finally, section 5 gives a conclusion of this paper.

## 2 Related Work

Face recognition research has got significant improvement over the past decades. Tolba *et al.* [25] and Zhao *et al.* [34] have made a comprehensive survey of face recognition methods. Early face recognition methods worked on holistic facial features. It has been proved that low-level feature descriptors could be an effective approach in face recognition [13,19,4,33,3]. Local descriptors such as LBP, SIFT, and HOG could extract distinctive text features. Luo *et al.* proposed to use the person-specific SIFT features and a simple non-statistical matching strategy

combined with the local and global similarity on key-points clusters to solve face recognition problems in [19]. In [12], a logistic discriminative approach which learned the metric from a set of labeled image pairs, and a nearest neighbor approach which computed the probability for two images were proposed to address face recognition of large data sets.

The use of weight-based models is common in classification problem [29]. Most weight-based methods use a set of weights for all instances. Raymer *et al.* proposed a global feature weighting algorithm in [22] and [21]. This algorithm optimizes a vector of weights that is used to scale the original features. Considering prototype weighting, Fernandez *et al.* proposed a local system used with a prototype-based classifier [10]. The weights are iteratively calculated after applying a local data normalization. In most feature weighting methods based on nearest neighbor rule, the distance between the sample feature and the neighbor is used as the weights. In this paper, we use correct frequency of a feature as its weight. This weights is more discriminative than those using distance. Frequent item mining (association rule learning) is a popular and well researched method for discovering interesting relations between variables in large databases [1,2,27]. Basura Fernando *et al.* [11] use this method in image classification and got significant performance.

Boiman *et al.* claimed that the quantization of local image descriptors and the computation of image-to-image distance lead to an inferior performance of nearest neighbor based image classification problems in [6]. Cheng et al. applied the image-to-class method to 3D hand gesture recognition [8]. In this paper, we also adopted image-to-class method too.

### 3 The Proposed Face Recognition Method

In this paper, We first use the image retrieval method to handle the large training data set, the images which are similar with the input image will be selected. Then we proposed online discriminative feature learning method, each feature or feature-pair will be given a corresponding weight. At last, we use k-nearest feature voting method to classify the face image.

#### 3.1 Feature Extraction

We propose to use parametric spatial-SIFT feature to represent a face image. The parametric spatial-SIFT  $f_i^s$  consists of the SIFT feature descriptor  $f_i$  and its relative location  $l_i = [r_x, r_y]$  with a parameter  $\alpha$ .

$$f_i^s = [f_i, \alpha l_i], 0 \leq \alpha \leq 1, \quad (1)$$

where  $l_i$  is the normalized location of the SIFT descriptor. The parameter  $\alpha$  is introduced to adjust the influence of the spatial constraint. In the experiment section, we examine the performance of face identification with a varied  $\alpha$ .

### 3.2 Similar Face Retrieval

The visual words in this paper are generated from the spatial-SIFT features by  $K - Means$  clustering method. Then we use the spatially-constrained similarity measure (*SCSM*) proposed in [23] to measure the similarity between query image and gallery image. Denote the query image by  $Q$ , and the spatial-SIFT features extracted from  $Q$  by  $\{f_1, f_2, \dots, f_m\}$ . Similarly, denote the gallery image by  $D$ , and the features in  $D$  by  $\{g_1, g_2, \dots, g_n\}$ . The similarity between  $Q$  and  $D$  is defined as:

$$S(Q, D) = \sum_{k=1}^N \sum_{\substack{(f_i, g_j) \\ f_i \in Q, g_j \in D \\ w(f_i) = w(g_j) = k \\ \|G(f_i) - G(g_j)\| < \varepsilon}} \frac{idf^2(k)}{tf_Q(k) \cdot tf_D(k)} \quad (2)$$

where  $k$  denotes the  $k$ -th visual word in the dictionary, and  $N$  is the vocabulary size.  $w(f_i) = w(g_j) = k$  means  $f_i$  and  $g_j$  are both assigned to visual word  $k$ .  $G(f) = (x_f, y_f)$  is the location of feature  $f$ . Thus the face images are normalized previously. The spatial constraint  $\|G(f_i) - G(g_j)\| < \varepsilon$  means that the locations of two matched features should be sufficiently close.  $idf(k)$  is the inverse document frequency of visual word  $k$ , and  $tf_Q(k)$  is the term frequency of visual word in  $Q$ . Similarly,  $tf_D(k)$  is the term frequency of visual word  $k$  in  $D$ .

Each face image in the database can get a similarity score. Then  $M$  face images with highest scores are selected as the similar face subset. The similar face subset will be used as training set in the recognition stage, which will be introduced in the next section.

### 3.3 Discriminative Feature Mining

**Single Feature Mining.** To learn the feature's discriminative ability, a feature's frequency in its class and other class are considered. The leave-one-out method is exploited to mine frequent feature and learn the features' weights in the similar face subset. Initially all features have the same weight  $w_i = w_0$ . If  $f_j$  is searched as the nearest neighbor of feature  $f_i$ , we increase the weight of  $f_j$  by  $\Delta w$  when  $f_j$  has the same label as the label of feature  $f_i$ , otherwise decrease the weight by  $\Delta w$  as shown in Eqn. (3), where  $n$  is the number of updates. It continues until all features in the query image find their neighbors in the training feature pool.

$$\begin{aligned} w^{n+1}(f_j) &= w^n(f_j)(1 + \Delta w), \text{ if } L(f_j) = L(f_i), \\ w^{n+1}(f_j) &= w^n(f_j)(1 - \Delta w), \text{ if } L(f_j) \neq L(f_i), \end{aligned} \quad (3)$$

where  $w^0(f_j) = w_0$ , and  $w_0$  is the initial weight for each feature. Function  $L(\cdot)$  takes a feature or a feature pair as input and returns its class label. Notation  $w^n(f_j)$  means the weight of feature  $f_j$  after its  $n^{th}$  nearest neighbor searching. The frequency and discriminative ability of each feature is thus quantized into a corresponding weight.

**Pair-Wise Feature Mining.** We further proposed to use the weighted frequent feature pairs as a new set of mid-level features to represent a face. First, we select  $p$  feature pairs from the single features with high weight with  $p \ll s$ , where  $s$  is the number of single features. Then, we learn the feature pairs' weights following the same procedure used in the single feature weight learning.

### 3.4 Feature Voting

We use the nearest neighbor voting as our recognition classifier. From Section 3.3, we have learned weights for both frequent single feature and feature pairs in the similar face subset. In the conventional voting based schemes, the nearest neighbor votes one score to its own class, which treats all features equally. In our process, the nearest subset feature contributes a weight directly to its corresponding class, which means that a test feature has a certain possibility to be in that class. By weighted feature voting, we can make sure that a feature with low discriminative ability makes less contribution to the final decision.

For a test face image  $I_Q$ , we first extract its spatial-SIFT descriptors as mentioned in Section 3.1. Then the image is represented by a set of spatial-SIFT features and  $N$  is the number of features in the test image as in Eqn. (4).

$$I_Q = [f_{Q_i}, i = 1, 2, \dots, N]. \quad (4)$$

For a test feature  $f_{Q_i}$ , by  $k$ -nearest neighbor searching, we find its nearest neighbor  $f_j$  in the training feature pool, and the nearest feature has a weight of  $w_j$ . Assume that there are  $K$  different face classes, and the face label of feature  $f_j$  is  $L_k$ . Then  $f_j$  votes  $w_j$  for the face label  $L_k$ . The rest of the test features will follow the same procedure to vote for their face labels.

As the same way in the single feature voting method, for a feature pair  $p_j$ , if both features in the pair are searched as  $k$ -nearest neighbors, we then votes the weight  $w_{p_j}$  of pair  $p_j$  for the pair's label. The final voting score is composed of single feature voting score as shown in Eqn. (5,6,7). As the pair-wise feature's weight is more confident than the single feature's weight, we add a parameter  $\lambda$  ( $\lambda > 1$ ) to the pair-wise feature's score.

$$S_{single} = \sum_j w_j, \quad L(f_j) = i, \quad (5)$$

$$S_{pair} = \sum_j w_{p_j}, \quad L(p_j) = i, \quad (6)$$

$$S_i = S_{single} + \lambda S_{pair}. \quad (7)$$

where the initial score of each class is 0,  $i = 1, \dots, K$ . At last, the test image  $Q$  is assigned to the class which has the highest voting score, defined as:

$$L(Q) = L(S_m), \quad \text{where } S_m = \max_i S_i \quad (8)$$

## 4 Experiment Results and Analysis

To validate the effectiveness of our method, experiments are carried out on two face data sets: the AR face data set and Pubfig83 face data set. We did experiments of disguised and occluded partial face recognition on the AR, and natural scenes face recognition on Pubfig83.

### 4.1 Data Sets

**AR Data Set.** The AR data set [20] contains 126 subjects, including 70 male and 56 female, respectively. For each subject, there are 26 face pictures taken in two different sessions (each session has 13 face images). In each session, there are 3 images with different illumination conditions, 4 images with different expressions, and 6 images with different facial disguises (3 images wearing sunglasses and 3 images wearing scarf, respectively)

**Pubfig83 Data Set.** The PubFig83 data set is a recently released face data set [16], which consists of a set of nearly 16,000 image that depicts 83 people, most of whom are well-known actors. Faces in Pubfig83 face data set vary highly from poses, illumination, occlusion and so on, so it is quite difficult to distinguish faces between each other. Obviously, Pubfig83 is much more cluttered, has more pose variations than AR, since it is from face images in the wild.

### 4.2 Experiment Settings

For the subjects in the Pubfig83 data set, we incrementally chose the identities from 2 to 83. To compare our method with *Xiong's* [31], we chose the first 90 images from each subject as training set, the first 78 images of each subject are using as gallery set, and the last 12 images are using as query set. For the gallery set, all images were normalized to  $128 \times 128$  pixels according to the facial points.

For the AR database, a subset containing 50 male subjects and 50 female subjects were selected from the first session in the AR data set as in [28]. For each identity, 14 images (without occlusion) were used for training, while 6 images with sunglasses and 6 images with scarves were selected for testing. For fair comparison with existing holistic methods, all these probe images and gallery images were cropped to  $128 \times 128$  pixels and properly aligned.

To determine the parameters of our method, we use cross-validation scheme to chose the best value for each parameter. First we define a set of values for each parameter, then we do cross-validation of each value. Finally, we can determine the value of each parameter, the parameter in spatial-SIFT  $\alpha = 0.55$  from set  $\{0, 0.05, \dots, 0.95, 1\}$ , the parameter for the  $k$ -nearest neighbor classifier  $k = 2$  from set  $\{1, 2, \dots, 6, 7\}$ , the parameter of discriminative feature learning  $\Delta w = 0.15$  from set  $\{0.05, 0.15, \dots, 0.85, 0.95\}$ . The other parameters we set based on experience, the initial value of  $w_0 = 1$ , feature pair parameter  $\lambda = 1.1$  and  $p = s/100$ .

### 4.3 Results and Analysis

**Experiment 1: Comparison with Different SIFT Features.** We conducted face recognition experiments on Pubfig83 data set. To demonstrate the effectiveness of our face retrieval and pair-wise feature learning method, we use nearest voting based on SIFT feature and spatial-SIFT feature as baseline method for comparison.

Table 3 shows the results of our method and the baseline methods, it is obviously that our proposed method achieved better performance. We can see our proposed similar face retrieval and discriminative feature learning method work effectively. Note that, the experiment are evaluated only on the first 20 subjects of Pubfig83 data set, using the same gallery images and query images as in section 4.2.

**Experiment 2: Face Recognition Under Disguise.** The AR face data set was selected to show our method’s effectiveness under occlusion data. Table 2 shows the recall and precision results of face retrieval by varying the number of similar face images. Table 1 records the recognition accuracy on the AR face data set with sunglasses, scarf and both, respectively. Our proposed method shows superior performance over the other state-of-the art methods on the AR data set, which could be credited to our discriminative feature learning scheme: the more discriminative feature to feature pair will have higher weight, and features from sunglass or scarf region will have lower weights. Thus the feature voting classifier will be robust to the noise and outliers.

**Table 1.** Results on AR data set

Method	Sunglass(%)	Scarf(%)	Sunglass + Scarf(%)
SRC [30]	87.5	59.50	73.25
CRC [32]	68.50	90.50	79.50
RoBM [24]	84.50	80.70	82.60
Stringfaces [7]	88.00	96.00	92.00
NNCW [17]	88.44	62.19	75.32
$l_1 - l_{struct}$ [15]	92.50	69.00	80.80
MLERPM [28]	98.00	97.00	97.50
Our method	97.50	98.83	98.16

**Table 2.** The similar face retrieval results on AR data set

Recall(%)	3.07	17.60	32.68	47.17	59.10	71.28	83.21	97.54
Precision(%)	81.50	71.40	68.70	65.20	63.02	51.17	47.32	37.20

**Experiment 3: Face Recognition Under Natural Scenes.** We compare the proposed method against several representative recognition methods on Pubfig83 dataset. The experimental results are listed in Table 3 which shows that our method is comparable with the "supervised AHR". In Table 4, LBP refers to directly using concatenated LBP features; Eigenfaces and Fisherfaces learn subspaces over concatenated LBP features; LBP, Eigenfaces, and Fisherfaces all use the cosine similarity for the final recognition.

**Table 3.** Results on SIFT features

Feature	Accuracy(%)
SIFT	79.5
Spatial-SIFT	81.0
Our method	86.5

**Table 4.** Results on Pubfig83

Method	Accuracy(%)
LBP [3]	56.3
Eigenfaces [26]	56.3
Fisherfaces [5]	60.2
SRC [30]	75.2
AHR [31]	85.5
Our method	79.52

## 5 Conclusions and Future Work

In this paper, we have proposed a face recognition method by using face retrieval and discriminative feature learning. We proposed to use spatial-SIFT instead of original SIFT features, then we find the similar faces based on the image retrieval method. After the similar face set are determined, we proposed to use discriminative feature learning for the features. Each feature and feature pair would be learned a corresponding weight. Features which are more discriminative would be given higher weights, noise features will be given lower weights, guaranteeing the robustness of our method. Experimental results on two widely used face data sets were presented to show the efficiency and limitations of our proposed method, our future work will be concerned to learning structure of faces to build high level features.

**Acknowledgement.** This work is supported by the grant ‘National Natural Science Foundation of China (NSFC)’ (No. 61273256 and No. NO.61305033).

## References

1. Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) *Knowledge Discovery in Databases*, pp. 229–248. AAAI Press (1991)
2. Agrawal, R., Ramakrishnan, Srikant, o.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)

3. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J.(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
4. Albiol, A., Monzo, D., Martin, A., Sastre, J., Albiol, A.: Face recognition using hog–ebgm. Pattern Recognition Letters 29(10), 1537–1543 (2008)
5. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on PAMI 19(7), 711–720 (1997)
6. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE CVPR (2008)
7. Chen, W., Gao, Y.: Recognizing partially occluded faces from a single sample per class using string-based matching. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 496–509. Springer, Heidelberg (2010)
8. Cheng, H., Dai, Z., Liu, Z.: Image-to-class dynamic time warping for 3d hand gesture recognition. In: IEEE ICME (2013)
9. Cheng, H., Yu, R., Liu, Z., Liu, Y.: A pyramid nearest neighbor search kernel for object categorization. In: IEEE ICPR (2012)
10. Fernández, F., Isasi, P.: Local feature weighting in nearest prototype classification. IEEE Trans. on Neural Networks 19(1), 40–53 (2008)
11. Fernando, B., Fromont, E., Tuytelaars, T.: Effective use of frequent itemset mining for image classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 214–227. Springer, Heidelberg (2012)
12. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: IEEE ICCV (2009)
13. Heisele, B., Serre, T., Poggio, T.: A component-based framework for face detection and identification. International Journal of Computer Vision 74(2), 167–181 (2007)
14. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
15. Jia, K., Chan, T.-H., Ma, Y.: Robust and practical face recognition via structured sparsity. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 331–344. Springer, Heidelberg (2012)
16. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: IEEE ICCV (2009)
17. Liu, Y., Wu, F., Zhang, Z., Zhuang, Y., Yan, S.: Sparse representation using non-negative curds and whey. In: IEEE CVPR (2010)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
19. Luo, J., Ma, Y., Takikawa, E., Lao, S., Kawade, M., Lu, B.-L.: Person-specific sift features for face recognition. In: IEEE ICASSP (2007)
20. Martinez, A.M.: The ar face database. CVC Technical Report 24 (1998)
21. Raymer, M.L., Doom, T.E., Kuhn, L.A., Punch, W.F.: Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics 33(5), 802–813 (2003)
22. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. IEEE Trans. on Evolutionary Computation 4(2), 164–171 (2000)

23. Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In: IEEE CVPR (2012)
24. Tang, Y., Salakhutdinov, R., Hinton, G.: Robust boltzmann machines for recognition and denoising. In: IEEE CVPR (2012)
25. Tolba, A., El-Baz, A., El-Harby, A.: Face recognition: A literature review. International Journal of Signal Processing 2(2) (2006)
26. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
27. Uno, T., Asai, T., Uchida, Y., Arimura, H.: Lcm: An efficient algorithm for enumerating frequent closed item sets. In: FIMI, vol. 90. Citeseer (2003)
28. Weng, R., Lu, J., Hu, J., Yang, G., Tan, Y.-P.: Robust feature set matching for partial face recognition. In: IEEE ICCV (2013)
29. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review 11(1-5), 273–314 (1997)
30. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. on PAMI 31(2), 210–227 (2009)
31. Xiong, Y., Liu, W., Zhao, D., Tang, X.: Face recognition via archetype hull ranking. In: IEEE ICCV (2013)
32. Zhang, D., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: IEEE ICCV (2011)
33. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In: IEEE ICCV (2005)
34. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. Acm Computing Surveys (CSUR) 35(4), 399–458 (2003)

# Single-Sample Face Recognition via Fusion Variant Dictionary

Ying Tai, Jian Yang, Jianjun Qian, and Yu Chen

Nanjing University of Science and Technology  
Nanjing 210094, P.R. China

{tyshiwo, chenyu1523}@gmail.com, csjyang@njust.edu.cn, qjjtx@126.com

**Abstract.** This paper presents a novel method called sparse representation based classification via fusion variant dictionary (FSRC) for single-sample face recognition. There are two points to be highlighted in our method: (1) A specific preprocessing step is introduced to help the gray level of the testing sample distributed uniformly. (2) A fusion variant dictionary is proposed including two parts: the first part is an intra-class variant term, which can help represent the moderate illuminations, expressions and disguises; the second part is a noise term, which can help remove the common noise (caused by pixel noise, severe illumination or our preprocessing step) in testing samples. Extensive experiments on public face databases demonstrate advantages of the proposed method over the state-of-the-art methods, especially in dealing with image corruption and severe illumination.

**Keywords:** Single-sample, face recognition, sparse representation, noise term.

## 1 Introduction

Face recognition is a classical problem in computer vision. In recent years, lots of practical systems have been developed, such as Google Picasa [7], face.com [8] and Face++ [9]. Since many applications on law enforcement, e-passport, driver license or identification card may only offer a single facial image per subject, the single-sample face recognition problem attracts researchers' attention.

Sparse representation based classification (SRC) [5] has shown its robustness to many problems in face recognition and gave some impressive results. However, it needs sufficient training images per subject. To overcome this drawback of SRC, several works have been done by researchers. Deng et al. [1] proposed an Extended SRC, which constructs an intra-class variant dictionary from a generic set to represent the variation between the testing and training samples. The author also proposed another method [2] to deal with the situation when the training samples are corrupted. In addition, Zhuang et al. [4] sought additional illumination examples of face images from other subjects to form an illumination dictionary for single-sample face alignment and recognition. More recently, Yang et al. [3] learned a sparse variation dictionary, jointly with an adaptive project

from the generic set to the training set, which helps to handle various variations in face images.

Although the methods mentioned above have shown their effectiveness, there's still a critical issue that needs to be solved. As we mentioned before, for single-sample face recognition, the intra-class variation between the test image and its relevant training image can be approximated by a sparse linear combination of the intra-class differences from sufficient number of generic faces [1, 2, 3, 4]. Here the intra-class variation includes various illuminations, expressions, poses and disguises. However, because of the randomness of the corrupted pixel's location, the intra-class variant dictionary cannot represent the unknown pixel noise in query images. As we know, Wright et al. [5] proposed a robust version of SRC (R-SRC) to increase its robustness to occlusion and corruption. By appending an identity matrix to the training set, the author solved the problem of calculating the sparse residual through seeking the sparse representation coefficients. This robust version, however, needs to construct a very large dictionary, which makes it time-consuming.

In this paper, we propose a novel method called sparse representation based classification via fusion variant dictionary (FSRC). Unlike R-SRC, which seeks the sparsest representation coefficient, our method gives an novel insight and seeks the sparsest residual straightway. In FSRC, except the intra-class variant part mentioned in [1], a novel noise term is introduced to supplement the variant dictionary along with a specific preprocessing step. The preprocessing step is designed to add an *unbalanced noise* over the query image, which makes its gray level distributed uniformly and we assume that the uniformly corrupted image contains a common noise, which can be removed by our noise term. Fig. 1 gives an intuitive illustration about the effect of our noise term. The noise term here is a vector with the same atoms. Without it, the residual caused by the intra-class variant part only may not be sparse; but with it, the residual generated adaptively in our model becomes sparse. Besides, as a derivation of ESRC, our method only introduce one additional item, which almost has no effect on the time cost. However, the robust version of ESRC (R-ESRC), which adds an identity matrix into the variant dictionary and is introduced to increase ESRC's robustness to corruption, suffers a lot because of its huge dictionary. Based on these observations, our method gains two advantages compared with those state-of-the-art methods: (1) shows impressive performance to various face variations, especially in dealing with large image corruption and severe illumination; (2) runs hundreds of times faster than the robust version along with a better accuracy. Experiments on public face databases demonstrate FSRC's advantages.

## 2 SRC via Fusion Variant Dictionary

This section introduces our fusion variant dictionary and a corresponding sparse representation based classification. Then, we give a discussion on our noise term and provide a specific preprocessing step.

## 2.1 Fusion Variant Dictionary

As we mentioned before, our fusion variant dictionary including an intra-class variant part  $V$  and a noise term  $b$ . Just like [1], the variant matrix  $V$  can be constructed in various ways from a generic set. Here, we choose two ways to generate it. Given a generic set  $G$ , which contains multiple images per subject. The  $s_i$  samples of subject  $i$  (stacked as vectors) form a matrix  $G_i \in R^{m \times s_i}, i = 1, \dots, t, \sum_{i=1}^t s_i = s$ . Here,  $m$  is the dimension of the image,  $t$  is the number of subjects and  $s$  is the total number of the images in  $G$ . According to whether there is a natural sample for each subject or not, we adopt different strategies to construct the intra-class variant part. If the natural sample exists, the variant bases can be generated by subtracting the natural image from other images of the same class:

$$V^{(1)} = [V_1^- - g_1^* e_1, \dots, V_t^- - g_t^* e_t] \in R^{m \times (s-t)} \quad (1)$$

where  $e_i = [1, \dots, 1] \in R^{1 \times (s_i-1)}$ ,  $g_i^*$  is the natural sample in class  $i$  and  $V_i^-$  is the reduced data matrix of class  $i$ . If the natural sample is not available, we build the variant part as follows:

$$V^{(2)} = [V_1 - c_1 e_1, \dots, V_t - c_t e_t] \in R^{m \times s} \quad (2)$$

where  $e_i = [1, \dots, 1] \in R^{1 \times s_i}$ ,  $c_i$  is the class centroid of class  $i$ . Apart from the intra-class variant part, the noise term in our model is represented as

$$b^T = [1, \dots, 1] \in R^{1 \times m}. \quad (3)$$

Therefore, our fusion variant dictionary can be constructed as

$$F^{(1)} = [V_1^- - g_1^* e_1, \dots, V_t^- - g_t^* e_t, b] \in R^{m \times (s-t+1)} \quad (4)$$

or

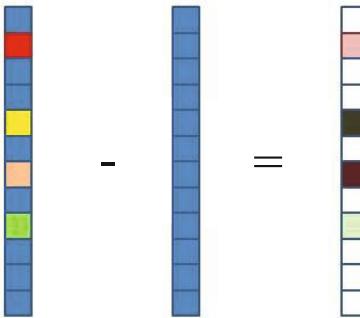
$$F^{(2)} = [V_1 - c_1 e_1, \dots, V_t - c_t e_t, b] \in R^{m \times (s+1)}. \quad (5)$$

## 2.2 SRC via Fusion Variant Dictionary

After we have a fusion variant dictionary, the single-sample face recognition problem can be seen as finding a sparse linear representation of the query image over a combination of the training set and the fusion variant dictionary. Given a training set  $A = [A_1, \dots, A_n] \in R^{m \times n}$  including images from  $n$  classes, a query image  $Y \in R^m$  and a fusion variant dictionary  $F \in R^{m \times k}$  (the dictionary size  $k$  depends on the generic set and construction method), then the problem can be formulated as

$$Y = A\alpha + F\beta + E \quad (6)$$

where  $\alpha$  is the representation coefficient regard to the training set,  $\beta$  is the representation coefficient regard to the fusion variant dictionary and  $E$  is the



**Fig. 1.** An intuitive illustration about the effect of the noise term in which different color means different pixel value. The left column represents an original signal. The middle column represents the noise term and the right column represents the sparse residual.

residual. If  $n$  is reasonably large, the coefficients in  $\alpha$  should be sparse. If sufficient variations are provided by the intra-class variant part, the coefficients in  $\beta$  should be sparse too. Therefore, we can recover the two sparse representation coefficients simultaneously by  $L_1$ -minimization. As we know, the  $L_1$ -minimization constraint on the representation coefficients brings the robustness to some variations [18]. Based on this situation, the noise term is designed to strengthen our method's ability to deal with pixel noise. Suppose  $\beta_b$  is the relevant coefficient with respect to the noise term  $b$ . The Eq. (6) can be rewritten as:  $E = Y - (A\alpha + F_{1:(k-1)}\beta_{1:(k-1)}) - b\beta_b$ . We cast this problem as finding the most suitable  $\beta_b$  to make the residual  $E$  sparse.

Based on the discussion mentioned above, we propose a corresponding sparse representation based classification via fusion variant dictionary (FSRC). The entire procedure is summarized in Algorithm 1. Here, we use the augmented Lagrange multipliers (ALM) method [10-14] to solve the  $L_1$ -minimization problem<sup>1</sup> with the parameter  $\lambda = 1$ .

### 2.3 A Specific Preprocessing Step

In this section, we give a discussion on the noise term and then propose a specific preprocessing step. As we mentioned before, the noise term here is a vector whose atoms are the same. Apparently, if there is a common noise in the query image, our noise term plays a role. However, sometimes the degree of corruption spans the query image unevenly or even differs greatly, which means some pixels may be heavily corrupted while the others seem to be clean. At this time, our model may fail because when we remove the common noise in the heavily corrupted pixels,

<sup>1</sup> This optimization method is described in detail in [14] to solve the  $L_1$ -minimization problem and its code can be downloaded at <http://www.eecs.berkeley.edu/~yang/software/l1benchmark/>.

**Algorithm 1.** SRC via Fusion Variant Dictionary

**Input:** A training set  $A = [A_1, \dots, A_n] \in R^{m \times n}$  including images  $A_1, \dots, A_n \in R^m$  from  $n$  classes, a query image  $Y \in R^m$ , a fusion dictionary  $F \in R^{m \times k}$  and a model parameter  $\lambda$ .

1: Normalize the columns of  $A$  and  $F$  to have unit L<sub>2</sub>-norm.

2: Solve the L<sub>1</sub>-minimization problem

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \arg \min \left\| [A, F] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - Y \right\|_1 + \lambda \left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_1, \quad (7)$$

where  $\alpha, \hat{\alpha} \in R^n$ ;  $\beta, \hat{\beta} \in R^k$ .

3: Compute the residuals

$$r_i(Y) = \left\| Y - [A, F] \begin{bmatrix} \delta_i(\hat{\alpha}) \\ \hat{\beta} \end{bmatrix} \right\|_1, i = 1, \dots, n \quad (8)$$

where  $\delta_i(\hat{\alpha}) \in R^n$  is a new vector whose only nonzero entries are the entries in  $\hat{\alpha}$  those are associated with class  $i$ .

**Output:**  $Identity(Y) = \arg \min_i r_i(Y)$

the clean pixels are also affected. To solve this problem, we introduce a specific preprocessing step to help the gray level of the query image distributed uniformly so as to create a common noise artificially. A suitable preprocessing step is very important for our noise term, here we conduct it through two operations. The first operation is gamma correction, which is conducted on the whole image. As we know, gamma correction is a nonlinear gray-level transformation [15]. It enhances the local dynamic range of the image in dark or shadowed regions while compressing it in bright regions and at highlights. The purpose of this operation is to add an *unbalanced noise* over the whole image, which eliminates the situation of uneven corruption. However, gamma correction can't handle the black pixels, which are very likely to be the noise. Therefore, the second operation is to remove the black pixels. Specifically, we adopt the way of average filtering to change the value of these pixels. We should note that if the query image is corrupted evenly over all of the pixels, the preprocessing step helps little. But if it is corrupted unevenly obviously, our preprocessing step helps a lot. Fig. 2 gives two examples from the extended Yale B database. We can see that the illumination variation spans evenly among the pixels in the example from Subset 3, while it differs a lot in the example from Subset 4. The processed images represent the effect of our preprocessing step. The one from Subset 3 changes little, but the other shows significant changes.

### 3 Experiments

In this section, we present several experiments to demonstrate the effectiveness of our method. Three public databases are used in our experiments, including the AR database [20], the Extended Yale B database [16, 17] and the CMU



**Fig. 2.** Two examples from the extended Yale B database. **Top:** The example from Subset 3. **Bottom:** The example from Subset 4. (a) The original images. (b) The processed images after our preprocessing step.

Multi-PIE database [6]. Here, we compare our proposed FSRC and the preprocessed version (Pre-FSRC) with some related work: ESRC [1] and SVDL [3]. For fair comparisons, we also give a preprocessed version of ESRC (Pre-ESRC) and SVDL (Pre-SVDL). However, since our preprocessing step has little impact on SVDL, we omit Pre-SVDL in our experiments. Meanwhile, the generic set is kept the same for FSRC, ESRC and SVDL.

### 3.1 Comparison with Related Work

The first experiment is a comparative test to severe intra-class variability, which is conducted on the AR database [20]. As shown in [1], we choose a random subset including 80 subject. For each subject, 13 images from Session 1 are selected: the image with natural expression and illumination for training, the others with severe intra-class variation for testing. The generic set is constructed by another 20 subjects (also with 13 images per subject) and our intra-class variation part is generated by Eq. (1). Therefore, our fusion variant dictionary contains 241 bases and images are cropped to  $100 \times 73$ . The error rates are shown in Table 1. What's more, Table 2 lists the specific error rate of every method under different variabilities introduced in [1]. From the results, We can see that: (1) SRC shows a poor performance for the lack of sufficient training samples per subject. (2) Because of the evenly noise distribution in the images, the preprocessing step helps little here. (3) Since the preprocessing step is designed for our method specifically, it may be unsuitable for other methods and lead to a worse effect. (4) Apart from illumination variation, FSRC outperforms ESRC comprehensively regardless of different expressions, disguises or the combination of illuminations and disguises. It should be noted that the reason for its failure in illumination may be the illumination variation in the AR database is inadequate. We will test our methods' robustness to specific problems in the following subsections.

**Table 1.** Comparative error rates (%) on the AR database

Methods	Err rates (%)
SRC [1]	40.31
ESRC [1]	10.63
Pre-ESRC	12.08
SVDL	11.46
FSRC	7.50
Pre-FSRC	<b>7.40</b>

**Table 2.** Comparative error rates (%) under different variabilities on the AR database

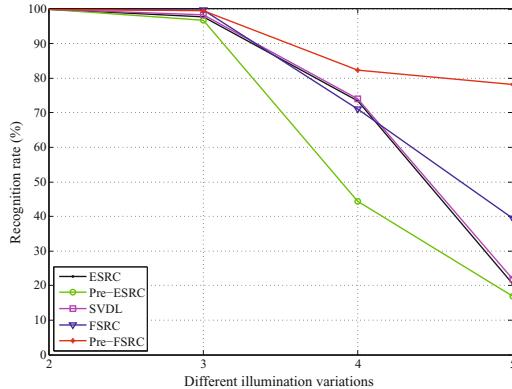
Err rates (%)	SRC [1]	ESRC [1]	Pre-ESRC	SVDL	FSRC	Pre-FSRC
Expression	17.9	10.8	11.7	12.1	<b>6.2</b>	6.7
Illumination	14.6	0.8	<b>0.4</b>	2.5	1.7	1.7
Disguise	48.1	10.6	11.9	8.1	7.5	<b>6.9</b>
Disguise + Illumination	71.9	17.8	21.3	19.4	12.8	<b>12.5</b>

### 3.2 Robustness to Illumination Variations

In this experiment, we evaluate FSRC’s robustness to adequate illumination variations on the Extended Yale B database [16, 17], which contains 38 subjects under 9 poses and 64 illumination conditions. The frontal face images in this database can be divided into 5 subsets according to different illumination conditions. Here, we choose a random subset including 30 subjects. For each subject, the first frontal image with natural illumination in Subset 1 is selected for training and the images in Subsets 2, 3, 4, 5 are used for testing respectively. As for the variant dictionary, the other 8 subjects (with 64 images per subject) are selected as the generic set to generate the intra-class variation part by Eq. (2). And the fusion variant dictionary here contains 543 bases. The images are cropped to  $48 \times 42$  and the results are shown in Fig. 3. From the discussion in Section 2.3, we know our method may fail to deal with the images in Subset 4 for its uneven noise distribution. However, after the preprocessing step, the effect of our noise term is exhibited obviously. In addition, since the preprocessing step is designed for our method specifically, it may be unsuitable for other methods, which can be seen from the performance of Pre-ESRC. We note that our proposed method achieves better recognition rates than other methods in all cases, especially in Subset 5 with at least 57% improvement. This observation demonstrates FSRC’s robustness to deal with adequate illumination variations.

### 3.3 Robustness to Different Pixel Noise

In this section, we evaluate FSRC’s robustness to different pixel noise on the CMU Multi-PIE database [6], which contains images of 337 subjects captured in four sessions with simultaneous variations in pose, expression and illumination.



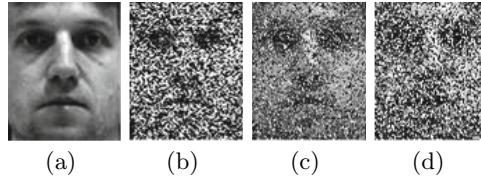
**Fig. 3.** Recognition rates (%) under different illumination variations on the Extended Yale B database

Here, we fix pose and expression, and choose a subset of this data set including the first 100 subjects from Session 1. For each subject, the frontal image with illumination 6 and neutral expression is selected for training and other 13 frontal images with various illuminations<sup>2</sup> and neutral expression are used for testing. The generic set here consists of another 8 subjects (with 14 images per subject from Session 1). The intra-class variation part is constructed by Eq. (2) and the fusion variant dictionary contains 113 bases. Besides, the images are cropped to  $100 \times 82$ . We conduct the tests on different pixel noise to show the effectiveness of FSRC. Specifically, salt and pepper noise [19], random corruption and speckle noise are introduced. Here, we fix the noise density of the first two kinds of noise as 50%, and the variance of speckle noise as 6. Fig. 4 gives an example of each pixel noise respectively. For all experiments, the location of noisy pixels is unknown to the algorithm. We introduce the robust version of ESRC in this experiment and the results of different methods are shown in Table 3. It can be seen that: (1) FSRC performs impressively when dealing with random corruption and salt and pepper noise, while fails to handle speckle noise. However, after the preprocessing step, our method does well in all cases and shows a big advantage over ESRC and SVDL. (2) R-ESRC really improves the ability of ESRC to deal with various pixel noises. However, compared with our method, it is still not good enough. What's more, our method runs hundreds of times faster than R-ESRC. We will show the comparison of running time between FSRC and R-ESRC in the following subsection.

### 3.4 Running Time

The last experiment is designed to compare the running time of FSRC and the robust version of ESRC, since running time is a very important factor in practical

<sup>2</sup> Illuminations {0,1,3,4,7,8,11,13,14,16,17,18,19}.



**Fig. 4.** The examples of different pixel noise. (a) The original image. (b) Image corrupted with 6 variance speckle noise. (c) Image corrupted with 50% random corruption. (d) Image corrupted with 50% salt and pepper noise.

**Table 3.** Recognition rates (%) under different pixel noise on the CMU Multi-PIE database

Recognition (%)	Speckle	Random corruption	Salt and pepper
ESRC	88.85	85.54	80.85
Pre-ESRC	88.15	83.23	83.77
R-ESRC	86.71	95.92	94.89
SVDL	91.82	90.39	86.20
FSRC	74.71	<b>98.31</b>	<b>99.23</b>
Pre-FSRC	<b>94.77</b>	<b>98.31</b>	98.62

face recognition systems. We conduct this experiment on the CMU Multi-PIE database with the same experimental setting as that in Section 3.3 except that the number of subjects is set as 50, 100, 150, 200 and 249, respectively. The desktop used is of 3.20 GHz CPU and with 7.68G RAM. Random corruption with noise density of 50% is used here and the average running time of FSRC and R-ESRC is listed in Table 4. As we can see, compared with FSRC, the running time of R-ESRC is much longer, over 3 minutes in all cases, while FSRC spends less than 3 seconds even when the subject number is 249. In general, our method runs hundreds of times faster than the robust version.

**Table 4.** The average running time (second) of FSRC and R-ESRC vs. subject number

	50	100	150	200	249
R-ESRC	181.91	191.62	197.77	225.58	258.28
FSRC	<b>0.32</b>	<b>0.68</b>	<b>1.04</b>	<b>2.15</b>	<b>2.73</b>

## 4 Conclusions

In this paper, we propose a novel single-sample face recognition method called sparse representation based classification via fusion variant dictionary (FSRC). In FSRC, a fusion variant dictionary including an intra-class variant part and a noise term is introduced to represent different variations between testing and

training samples. To make full use of the noise term, we provide a specific pre-processing step to help the gray level of the testing sample distributed uniformly. The extensive experiments clearly demonstrate that FSRC performs better than state-of-the-art single-sample face recognition methods along with impressive speed compared to the robust version.

## References

1. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE PAMI* 34, 1864–1870 (2012)
2. Deng, W., Hu, J., Guo, J.: In defense of sparsity based face recognition. In: *CVPR* (2013)
3. Yang, M., Gool, L.V., Zhang, L.: Sparse variation dictionary learning for face recognition with a single training sample per person. In: *ICCV* (2013)
4. Zhuang, L., Yang, A., Zhou, Z., Sastry, S., Ma, Y.: Single-sample face recognition with image corruption and misalignment via sparse illumination transfer. In: *CVPR* (2013)
5. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse re-presentation. *IEEE PAMI* 31, 210–227 (2009)
6. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. *Image and Vision Computing* 28, 807–813 (2010)
7. <http://picasa.google.com/> (accessed May 23, 2014)
8. <http://face.com/> (accessed May 23, 2014)
9. <http://cn.faceplusplus.com/uc/> (accessed May 23, 2014)
10. Lin, Z., et al.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055v2
11. Bertsekas, D.P.: Nonlinear programming. Athena Scientific (2004)
12. Cands, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *Journal of the ACM* 58(3), Article 11 (2011)
13. Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Fast l1-minimization algorithms and an application in robust face recognition: A review. In: *ICIP* (2010)
14. Yang, A., Ganesh, A., Zhou, Z., Sastry, S., Ma, Y.: Fast l1-minimization algorithms for robust face recognition. arXiv:1007.3753v4 (2012)
15. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *TIP* (2010)
16. Lee, K., Ho, J., Driegman, D.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE PAMI* 27(5), 684–698 (2005)
17. Georgiades, Belhumeur, P., Kriegman, D.: From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI* 23(6), 643–660 (2001)
18. Yang, J., Zhang, L., Xu, Y., Yang, J.: Beyond sparsity: the role of l1-optimizer in pattern classification. *Pattern Recognition* 45, 1104–1118 (2012)
19. Gonzalez, R., Woods, R.: Digital image processing. Pearson Prentice Hall (2007)
20. Martinez, A., Benavente, R.: The AR face database, CVC Technical Report (1998)

# Supervised Kernel Construction for Unsupervised PCA on Face Recognition

Yang Zhao, Wen-Sheng Chen<sup>\*</sup>, Binbin Pan, and Bo Chen

College of Mathematics and Computational Science, Shenzhen University  
Shenzhen Key Laboratory of Media Security  
Shenzhen, 518060, China  
[chenws@szu.edu.cn](mailto:chenws@szu.edu.cn)

**Abstract.** This paper aims to establish a novel framework for high-performance Mercer kernel construction. Based on a given kernel matrix incorporated the class label information, a nonlinear mapping is firstly generated and well-defined on the training samples. The partial data-defined mapping can be extended and well-defined on the entire pattern space by means of interpolatory technology. The analytic expression of the nonlinear mapping is then obtained. It theoretically shows that the function  $K(x, y)$ , created by the inner product of the nonlinear mapping, is a supervised Mercer kernel function. Our supervised kernel is successfully applied to unsupervised principal component analysis (PCA) method for face recognition. Two face databases, namely ORL and FERET databases, are selected for evaluations. Compared with KPCA with RBF kernel (RBF-PCA) method, experimental results demonstrate that KPCA with our supervised kernel (SK-PCA) has superior performance.

**Keywords:** Face Recognition, Supervised Mercer Kernel, Kernel PCA.

## 1 Introduction

Over the past decades, face recognition has become one of the most challenging technologies in the area of pattern recognition and computer vision because of variations of facial images, such as pose and illumination variations. These variations incur that the distribution of the facial data in original feature space is very complicated and usually nonlinear. So, the linear feature extraction methods, say PCA [1] and LDA [2], cannot achieve satisfactory performance. Kernel method is an effective means to tackle the nonlinear problem of face recognition [3]-[10]. The basic idea of kernel method is to find a nonlinear mapping  $\Phi$  which maps the input samples into a high dimensional feature space  $F$ , and then performs a linear classifier in the kernel feature space  $F$ . However, it is very difficult to learn a nonlinear mapping and the dimensionality of  $F$  is also large and perhaps infinite. Thereby, direct computation in  $F$  is infeasible. Fortunately, the linear feature extraction methods conducted in  $F$  just need to calculate the inner

---

\* Corresponding author.

product  $\langle \Phi(x), \Phi(y) \rangle_F$ . Based on Mercer kernel theory [12], the inner product  $\langle \Phi(x), \Phi(y) \rangle_F$  can be replaced with a kernel function  $K(x, y)$ , where  $x$  and  $y$  are two samples from the input feature space. This kernel trick allows us to execute the kernel method in kernel space  $F$  without knowing the nonlinear mapping. The kernel matrix, which is a symmetric and positive semi-definite matrix, plays an important role in kernel based machine learning. This paper will discuss how to construct a kernel matrix using the training data and further design a high-performance kernel function. It is known that the kernel matrices determined by the popular kernel functions, such as linear, polynomial and RBF kernels, do not make use of the class label information. The performances of these unsupervised kernel based learning methods will be degraded. To overcome this limitation, we previously proposed a kernel construction method using Lagrange interpolation strategy [8]. But this method is numerically unstable because of the Runge phenomenon caused by Lagrange interpolation [11].

To remedy the drawbacks of existing kernel learning algorithms, this paper proposes a new framework to design the supervised Mercer kernel with high-performance. As we know, kernel matrix associated with a certain kernel function records the similarity scores among the training samples. It is desired for high-performance kernel construction that the similarities among the intra-data have large scores, while the inter-data possess small similarities. To this end, the class label information is utilized in this paper to model a supervised kernel matrix, which is a block diagonal matrix generated using a radial basis function. Unlike Lagrange interpolation, we propose a methodology to construct some new interpolatory basis functions such that their values range in the interval  $[0, 1]$ . So, our interpolatory strategy can automatically eliminate the Runge phenomenon since the proposed method avoids the values of interpolatory basis functions from growing unboundedly. Based on the supervised kernel matrix and the new interpolatory strategy, this paper obtains the analytic expression of the nonlinear mapping and theoretically proves that the function  $K(x, y)$ , defined by the inner product of the nonlinear mapping, is a supervised Mercer kernel function. The supervised kernel (SK) is tested using kernel PCA approach for face recognition. Two publicly available face databases, namely ORL and FERET, are chosen for performance evaluations. Experimental results show that KPCA with our supervised kernel (SK-PCA) surpasses KPCA with RBF kernel (RBF-PCA).

The rest of this paper is organized as follows. Section 2 briefly introduces the related work. Section 3 gives the theoretic discussions on interpolatory basis functions and the supervised Mercer kernel constructions. Section 4 develops our SK-PCA algorithm. Experimental results are reported in section 5. Finally, section 6 draws the conclusions.

## 2 Related Work

This section will briefly introduce some related work such as PCA and Kernel PCA.

## 2.1 PCA

PCA is an unsupervised linear feature extraction and dimensionality reduction method, which finds the orthogonal linear transformation such that the mapped data along the principal component directions have the largest variances. In face recognition, PCA discovers the principal facial features which are called eigenfaces [1]. The facial images can be linearly expressed by the eigenfaces. However, PCA is a linear method which cannot uncover the underlying nonlinear structure of the facial images. Moreover, as an unsupervised method, PCA is used in a label-independent manner and thus not suitable for classification tasks.

## 2.2 Kernel PCA

To solve nonlinear problems, the classic PCA has been generalized to its kernel version, namely Kernel PCA (KPCA) [3,4]. In this method, the data from the input space are mapped into a high dimensional kernel space using a nonlinear mapping, where different classes of objects are supposed to be linearly separable. And then the classic PCA can be performed in the high dimensional feature space using kernel trick. However, the kernel matrices, which are computed by the commonly used kernels on the training data, are full matrices and cannot reflect the class label information. So, KPCA with the commonly used kernels, such as linear, polynomial and RBF kernels, is still an unsupervised learning approach. The accuracy of unsupervised KPCA will be affected in face recognition.

In the following sections, this paper discusses how to design a supervised Mercer kernel with high-performance.

## 3 The Proposed Method

This section proposes a theoretical framework on supervised Mercer kernel construction. Details are below.

### 3.1 Some Notations

Let  $d$  be the dimension of original feature space and  $C$  be the number of sample classes. The total original sample set  $X = \bigcup_{i=1}^C X_i$ , where the  $i$ th class  $X_i = \{x_j^i\}_{j=1}^{N_i}$  which contains  $N_i$  training samples.  $N (= \sum_{i=1}^C N_i)$  is the number of total training samples. Assume  $\Phi(x) : x \in R^d \rightarrow \Phi(x) \in F$  is the kernel nonlinear mapping, where  $F$  is the mapped feature space with dimension  $df$  ( $= \dim F$ ). The total mapped sample set is  $\Phi(X) = \bigcup_{i=1}^C \Phi(X_i)$ , and the  $i$ th mapped class is  $\Phi(X_i) = \{\Phi(x_j^i)\}_{j=1}^{N_i}$ . If  $K(x, y)$  is a Mercer kernel defined on  $R^d \times R^d$ , then there exists a nonlinear mapping  $\Phi$ , such that  $K(x, y) = \langle \Phi(x), \Phi(y) \rangle_F$ . We denote RBF kernel  $K_{RBF}(x, y)$  by  $K_{RBF}(x, y) = \exp(-\frac{\|x-y\|^2}{t})$  with  $t > 0$ .

### 3.2 Basis Function Construction

We construct  $N$  interpolatory basis functions  $L_j^i(x)$  as follows

$$L_j^i(x) = \frac{\omega_j^i(x)}{\sum_{J=1}^{N_I} \sum_{I=1}^C \omega_J^I(x)}, \quad (1)$$

where  $\omega_j^i(x) = \prod_{(p,q) \neq (i,j)} \|x - x_q^p\|$ ,  $q = 1, 2, \dots, N_p$ ,  $p = 1, 2, \dots, C$ ,  $x \in R^d$ . It can be easily verified that  $\omega_j^i(x)$  has the following property:

$$\omega_j^i(x) = \begin{cases} \prod_{(p,q) \neq (i,j)} \|x_j^i - x_q^p\|, & x = x_j^i \\ 0, & x \in X \setminus \{x_j^i\}. \end{cases}$$

Therefore, the interpolatory basis function  $L_j^i(x)$  satisfies that:

$$L_j^i(x_q^p) = \begin{cases} 1, & (p, q) = (i, j) \\ 0, & (p, q) \neq (i, j) \end{cases}, \text{ for all } x_q^p \in X.$$

From (1), we can see that basis function  $L_j^i(x)$  is a bounded function and ranges in the internal  $[0, 1]$ . For convenience, all basis functions are formed as a interpolatory basis vector function  $L(x)$  denoted by

$$L(x) = [L_1^1(x), \dots, L_{N_1}^1(x) | \dots | L_1^C(x), \dots, L_{N_C}^C(x)]^T. \quad (2)$$

### 3.3 Supervised Mercer Kernel Construction

Assume matrices  $K_i = (k_{jk}^{(i)})_{N_i \times N_i}$ , where  $k_{jk}^{(i)} = K_{RBF}(x_j^i, x_k^i)$ ,  $i = 1, 2, \dots, C$ . Let block diagonal matrix  $K$  be

$$K = \text{diag}(K_1, K_2, \dots, K_C) \in R^{N \times N}, \quad (3)$$

then  $K$  is a symmetric and positive semi-definite matrix, which is able to serve as a kernel matrix. It can be seen from matrix  $K$  that the similarities of the intra-data are large and that of the inter-data are small. Thus,  $K$  encodes the class label information of the training data.

By performing eigenvalue decomposition on  $K_i$ , we have  $K_i = \tilde{U}_i^T \Lambda_i \tilde{U}_i$ , where  $\tilde{U}_i$  is a  $N_i \times N_i$  orthogonal matrix and  $\Lambda_i$  is a diagonal matrix with non-negative diagonal entries. Let  $U_i = \Lambda_i^{\frac{1}{2}} \tilde{U}_i$  and  $U = \text{diag}(U_1, U_2, \dots, U_C) \in R^{N \times N}$ , matrix  $K$  has the decomposition  $K = U^T U \in R^{N \times N}$ , where

$$U = [u_1^1, \dots, u_{N_1}^1 | u_1^2, \dots, u_{N_2}^2 | \dots | u_1^C, \dots, u_{N_C}^C] \quad (4)$$

and  $u_j^i$  is the  $j + \sum_{k=1}^{i-1} N_k$  column of matrix  $U$ . A nonlinear mapping  $\Phi$  is defined on the training data set  $X$  by:

$$\Phi(x_j^i) = u_j^i, \quad j = 1, 2, \dots, N_i, i = 1, 2, \dots, C.$$

Above  $\Phi$  can be extended and well-defined on the whole input feature space by using interpolatory technique. In details, we extend the nonlinear mapping  $\Phi$  to the entire input feature space  $R^d$  as follows:

$$\Phi(x) = U \cdot L(x), \quad x \in R^d \quad (5)$$

where  $L(x)$  is the interpolatory basis vector function defined by (2) and matrix  $U$  is determined by (4). Then, we denote function  $K(x, y)$  by:

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_F,$$

where  $\langle \cdot, \cdot \rangle_F$  is a inner product in feature space  $F$ . It can be directly derived that

$$K(x, y) = L^T(x)KL(y). \quad (6)$$

In order to show that the function  $K(x, y)$  defined by (6) is a Mercer kernel, we need the following lemma.

**Lemma 1.** [12] If  $K(x, y)$  is a symmetric function defined on  $R^d \times R^d$ , and for any finite data set, it always yields a symmetric and positive semi-definite matrix  $K = (k_{ij})_{m \times m}$ , where  $k_{ij} = k(y_i, y_j)$ ,  $i, j = 1, 2, \dots, m$ , then function  $K(x, y)$  is a Mercer kernel function.

**Theorem 1.** Function  $K(x, y) = L^T(x)KL(y)$ , defined on  $R^d \times R^d$ , is a Mercer kernel function, where  $L(\cdot)$  and  $K$  are determined by (2) and (3) respectively.

*Proof.* Since function  $K(x, y)$  is apparently a symmetric function, it is merely to show that the Gram matrix  $G$  generated by  $K(x, y)$  on any finite training data is a positive semi-definite matrix. For any finite training data set  $\{y_l | l = 1, 2, \dots, n\} \subset R^d$ , the Gram matrix  $G$  can be calculated as  $G = [K(y_l, y_s)]_{n \times n}$ . If we denote matrix  $L_n$  by  $L_n = [L(y_1), L(y_2), \dots, L(y_n)]_{N \times n}$ , then Gram matrix  $G$  can be rewritten as  $G = L_n^T K L_n$ . For any column vector  $\alpha \in R^n$ , we have  $\alpha^T G \alpha = (L_n \alpha)^T K (L_n \alpha) \geq 0$  because  $K$  is a positive semi-definite matrix. It indicates that  $G$  is a symmetric and positive semi-definite matrix. The theorem is concluded from Lemma 1 immediately.

The constructed kernel  $K(x, y)$  has adopted the class label information and thus becomes a supervised kernel (SK) function, which will be evaluated using kernel PCA method.

## 4 Algorithm

This section develops a KPCA algorithm using SK function. Details are as follows.

**Step 1:** Construct symmetric and positive semi-definite matrix  $K = \text{diag}(K_1, \dots, K_C) \in R^{N \times N}$ , where  $K_i = [K_{RBF}(x_j^i, x_k^i)]_{N_i \times N_i}$  for  $x_j^i, x_k^i \in X_i$ .

**Step 2:** Let  $L(x) = [L_j^i(x)] \in R^{N \times 1}$ , where  $L_j^i(x)$  are the interpolatory basis functions defined by

$$L_j^i(x) = \frac{\omega_j^i(x)}{\sum_{J=1}^{N_I} \sum_{I=1}^C \omega_J^I(x)},$$

where  $\omega_j^i(x) = \prod_{(p,q) \neq (i,j)} \|x - x_q^p\|$ ,  $q = 1, 2, \dots, N_p$ ,  $p = 1, 2, \dots, C$ ,  $x_q^p \in X_p$ .

**Step 3:** The supervised kernel is constructed as

$$K(x, y) = L^T(x)KL(y).$$

**Step 4:** KPCA [4] with SK is performed for face recognition.

## 5 Experimental Results

This section will evaluate the performance of the proposed kernel based KPCA method for face recognition. In our experiments, linear PCA (PCA) [1] (as a benchmark here), PCA with RBF kernel (RBF-PCA) [4] and our method (SK-PCA) are chosen for comparisons on two datasets, namely ORL dataset and FERET dataset. The parameter  $t$  in RBF kernel is fixed to 1e3.

### 5.1 Facial Image Datasets

The ORL database contains 400 images of 40 persons and each person consists of 10 images with different facial expressions (open or closed eyes, smiling or not smiling), small variations in scales and orientations. The resolution of each image is  $112 \times 92$ , and with 256 gray levels per pixel. Facial image variations of one person from ORL database are shown in Figure 1.

For FERET database, we select 120 people, 6 images from each person. The resolution of each facial is also  $112 \times 92$ . FERET dataset is more challenging than ORL dataset since the variations in FERET database include pose, illumination, facial expression and aging. Images from two individuals are shown in Figure 2.



**Fig. 1.** Images of one person from ORL database



**Fig. 2.** Images of two persons from FERET database

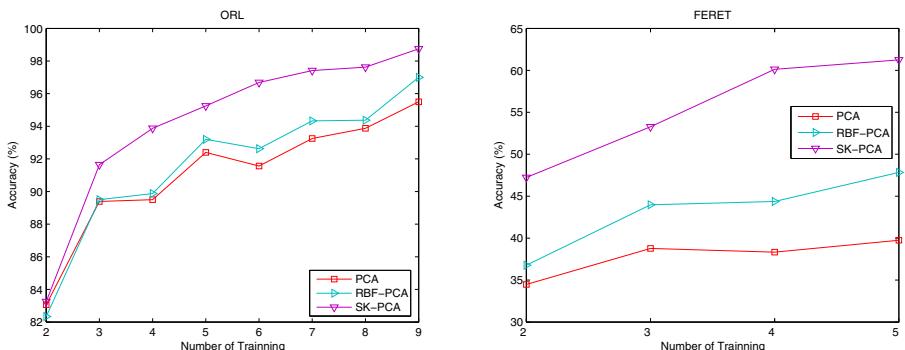
## 5.2 Comparisons on ORL Dataset

The experimental setting is as follows. The  $n(n = 2, 3, \dots, 9)$  images are randomly selected from each person for training, and the remaining  $(10 - n)$  images of each individual are for testing. The experiments are repeated 10 times and the mean accuracies are recorded in Table 1 and plotted in Figure 3 (left) respectively. It can be seen that the recognition rate of SK-PCA increases from 83.34% with training number 2 to 98.75% with training number 9, while the recognition rates of PCA and RBF-PCA increase from 83.06% and 83.25% with training number 2 to 95.50% and 97.00% with training number 9 respectively.

Experimental results show that the proposed SK-PCA method gives the best performance on ORL dataset.

**Table 1.** Mean accuracy(%) versus Training Number (TN) on ORL database

TN	2	3	4	5	6	7	8	9
PCA	83.06	89.39	89.50	92.40	91.56	93.25	93.88	95.50
RBF-PCA	83.25	89.50	89.88	93.20	92.63	94.33	94.37	97.00
SK-PCA	<b>83.34</b>	<b>91.64</b>	<b>93.88</b>	<b>95.25</b>	<b>96.69</b>	<b>97.42</b>	<b>97.63</b>	<b>98.75</b>



**Fig. 3.** Recognition rate on ORL face database (left) and FERET face database (right)

### 5.3 Comparisons on FERET Dataset

We randomly choose  $n(n = 2, 3, \dots, 5)$  images from each people for training, while the rest ( $6 - n$ ) images of each individual are selected for testing. The experiments are also run 10 times and the average accuracies are tabulated in Table 2 and plotted in Figure 3 (right) respectively. It can be seen that the recognition rate of SK-PCA increases from 47.23% (TN=2) to 61.25% (TN=5). In contrast, the accuracy of PCA ascends from 34.48% (TN=2) to 39.75% (TN=5), while the accuracy of RBF-PCA increases from 36.77% (TN=2) to 47.83% (TN=5).

Compared with PCA and RBF-PCA, our SK-PCA gives around 17.64% and 12.23% entire mean accuracy improvements on FERET dataset, respectively.

**Table 2.** Mean accuracy (%) versus Training Number on FERET database

TN	2	3	4	5
PCA	34.48	38.78	38.33	39.75
RBF-PCA	36.77	43.97	44.37	47.83
SK-PCA	<b>47.23</b>	<b>53.28</b>	<b>60.13</b>	<b>61.25</b>

## 6 Conclusions

In this paper, we propose a novel methodology to construct supervised kernel function with high-performance. The class label information is incorporated into the kernel matrix and the interpolatory basis functions, range in [0 1], are established to obtain the analytic expression of nonlinear mapping. The kernel function, generated using the inner product of nonlinear mapping, is theoretically proven to be a supervised Mercer kernel. The constructed supervised kernel is tested using kernel PCA for face recognition. Experimental results on ORL and FERET databases show that our SK-PCA method surpasses linear PCA and RBF-PCA methods. Especially, our supervised kernel can be applied to all the kernel based machine learning tasks.

**Acknowledgements.** This paper is partially supported by NSF of China Grant (61272252) and Science & Technology Planning Project of Shenzhen City (JCYJ20130326111024546). We would like to thank Olivetti Research Laboratory and Amy Research Laboratory for providing the face image databases.

## References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3, 71–86 (1991)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 711–720 (1997)

3. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* 10, 1299–1319 (1998)
4. Kim, K., Jung, K., Kim, H.J.: Face Recognition Using Kernel Principal Component Analysis. *IEEE Signal Processing Letters* 9, 40–42 (2002)
5. Yang, J., Jin, Z., Yang, J.Y., Zhang, D., Frangi, A.F.: Essence of Kernel Fisher Discriminant: KPCA plus LDA. *Pattern Recognition* 37, 2097–2100 (2004)
6. Eftekhari, A., Forouzanfar, M., Abrishami Moghaddam, H., Alirezaie, J.: Block-Wise 2D Kernel PCA/LDA for Face Recognition. *Information Processing Letters* 110, 761–766 (2010)
7. Ebied, R.M.: Feature Extraction Using PCA and Kernel-PCA for Face Recognition. In: 8th International Conference on Informatics and Systems, pp. 72–77. IEEE Press, New York (2012)
8. Chen, W.S., Yuen, P.C.: Interpolatory Mercer Kernel Construction for Kernel Direct LDA on Face Recognition. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 857–860. IEEE Press, New York (2009)
9. Chu, W.S., Chen, J.C., James, L.J.J.: Kernel Discriminant Transformation for Image Set-Based Face Recognition. *Pattern Recognition* 44, 1567–1580 (2011)
10. Chan, C.H., Tahir, M.A., Kittler, J., Pietikäinen, M.: Multiscale Local Phase Quantization for Robust Component-Based Face Recognition Using Kernel Fusion of Multiple Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1164–1177 (2013)
11. Süli, E., Mayers, D.F.: An Introduction to Numerical Analysis. Cambridge University Press, Cambridge (2003)
12. Schölkopf, B., Smola, A.J.: Learning with Kernels-Support Vector Machine, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2002)

# Medical Image Clustering Based on Improved Particle Swarm Optimization and Expectation Maximization Algorithm

Zheng Tang<sup>1,\*</sup>, Yuqing Song<sup>1</sup>, and Zhe Liu<sup>1,2</sup>

<sup>1</sup> School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

<sup>2</sup> School of Computer Science, Jilin Normal University, Siping, China  
tangzheng1119@163.com, yqsong@ujs.edu.cn, lxxc1016@gmail.com

**Abstract.** We proposed a hybrid clustering algorithm based on the improved particle swarm optimization algorithm and EM clustering algorithm to overcome the shortcomings of EM algorithm, which is sensitive to initial value and easy to sink into local minimum. First, get the optimal clustering number of any dataset to obtain the initial parameter of mixed model with the improved PSO algorithm, whose inertia weight increased and decreased along the fold line automatically. Then build the mixed density model of image data by multiple iterations of the EM algorithm. Finally divide all the pixel value of the image into corresponding branch of hybrid model with the Bayesian criterion to get the classification of image data. The proposed algorithm can increase the diversity of EM clustering algorithm initialization and promote optimization search in the global scope. Experimental results of simulation prove its accuracy and validity.

**Keywords:** Particle Swarm Optimization, EM algorithm, inertia weight, medical image clustering.

## 1 Introduction

As the medical images are widely used in clinical diagnosis in recently years, the status of medical image segmentation technique is increasingly important position in assisted the doctor diagnosis and treatment diseases, etc. It has the very important significance to the development of modern medicine, as the necessary means of a particular organization measure, diseased tissue extracting as well as three-dimensional reconstruction. Therefore, the research of medical image segmentation is of important theoretical value and broad application prospects.

Clustering, as one of the most effective image segmentation methods, refers to the act of partitioning an unlabeled dataset into groups of similar objects. Each group, known as a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In the past few decades, cluster analysis

---

\* Corresponding author.

has played a central role in a variety of fields ranging from engineering, computer sciences, life and medical sciences, to earth sciences, social sciences and some other fields [1]. As one of the most important area-based image segmentation methods, the scholars have developed a lot of typical clustering algorithms, such as K-means [2], fuzzy C-means [3], kernel density clustering [4] and finite mixture model [5].

The Expectation Maximization (EM) algorithm based on the Gaussian mixture models [6] is one of the general clustering methods which develop rapidly in recent years. In this algorithm, the EM algorithm was used to set up the parameters to be evaluated of the mixture model. Then the posterior probability of Bayesian criteria was used to partition the sample data into corresponding branch of the hybrid model. The clustering method is a semi-parametric density estimation method, which combines the advantages of parameter estimation and nonparametric estimation. It is not limited to a specific form of the probability density function and the complexities of the model only depends on the sample collection of the problem to be solved. The simulations reveal that our proposed algorithm achieved good results in image clustering problem. However, the EM algorithm has some obvious deficiencies, such as sensitivity to initial value and easy to fall into local minima, etc. A lot of improved EM algorithm variants were put out to overcome these problems. Ueda [6] proposed the Split and merge EM algorithm (SMEM) and Genetic-based EM algorithm (GAEM). Verbeek [7] proposed a Greedy EM algorithm (GEM) according to the incremental clustering implemented in greedy learning of Gaussian model. Zhao [8] proposed a random swap EM algorithm (RSEM) consists of removal and addition operations. Volodymyrz [9] acquired the initial mean vector by choosing a higher concentration of the points in the neighborhood space, and employed the truncated normal distribution to estimate the discrete matrix preliminary. But the clustering number of these algorithms is difficult to determine in advance and the parameter settings are complicated in all of these EM variants.

Particle Swarm Optimization (PSO) algorithm [10] is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling. The global space-based search of the PSO algorithm weakens the impact of the initial situation, and gets rid the tendency of the standard EM algorithm to get stuck in a local maximum. Then we proposed a hybrid PSOEM clustering algorithm based on the improved particle swarm algorithm to automatically find the clustering centers and get the parameter estimation of the mixed density model by multiple iterations of the EM algorithm, and finally realize the image clustering with Bayesian criteria.

## 2 EM Algorithm Based on the Gaussian Mixture Models

### 2.1 Gaussian Finite Mixture Models

Let  $Y = [Y_1, Y_2, \dots, Y_d]^T$  be a  $d$ -dimensional random variable, with  $y = [y_1, y_2, \dots, y_d]^T$  represents one particular outcome of  $Y$ . It is said that  $Y$  follows

a  $k$ -component finite mixture distribution if its probability density function can be written as:

$$p(y|\theta) = \sum_{m=1}^k \alpha_m p(y|\theta_m) \quad (1)$$

where  $\alpha_1, \alpha_2, \dots, \alpha_k$  are the mixing probabilities, each  $\theta_m$  is the set of parameters defining the  $m$ th component, and  $\theta_m \equiv \{\theta_1, \dots, \theta_k, \alpha_1, \dots, \alpha_k\}$  is the complete set of parameters needed to specify the mixture. And the  $\alpha_m$  must satisfy:

$$\alpha_m \geq 0, \sum_{m=1}^k \alpha_m = 1, m = 1, 2, \dots, k \quad (2)$$

We assume that all the components have the same function form(for instance, they are  $d$ -variant Gaussian), each one being characterized by the parameter vector  $\theta_m$ . Given a set of  $n$  independent and identically distributed samples  $y = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$ , and the log-likelihood corresponding to a  $k$ -component mixture is

$$\log p(y|\theta) = \log \prod_{i=1}^n p(y^{(i)}|\theta) = \sum_{i=1}^n \log \sum_{m=1}^k \alpha_m p(y^{(i)}|\theta) \quad (3)$$

It is well-known that the maximum likelihood (ML) estimation  $\hat{\theta}_{ML} = \arg \max_{\theta} \{\log p(y|\theta)\}$  cannot be found analytically. But the Bayesian maximum a posteriori (MAP) criterion  $\hat{\theta}_{MAP} = \arg \max_{\theta} \{\log p(y|\theta) + \log p(\theta)\}$  presented some prior  $p(\theta)$ on the parameters enable easier estimation of the maximum posterior probability. Whats more, the maximizations defining of the ML or MAP estimates should be under the constraint of formula (2).

## 2.2 EM Algorithm Based on the Gaussian Mixture Models

The method using maximum likelihood estimation (ML) or maximum a posteriori (MAP) for parameter estimation [11] is commonly known as EM algorithm, which is an iterative procedure to get the local maxima of  $\log p(y|\theta)$  or  $\{\log p(y|\theta) + \log p(\theta)\}$ . For the multi-dimensional Gaussian mixtures, let  $\theta = (\alpha_1, \mu_1, \Sigma_1, \alpha_2, \mu_2, \Sigma_2, \dots, \alpha_k, \mu_k, \Sigma_k)$  be the parameter of each Gaussian density function, where  $\mu_k$  is the mean value,  $\Sigma_k$  is the covariance matrix,  $\alpha_k$  is the proportion of each sample in the total samples. The maximum likelihood estimation with EM algorithm [12] is as follows:

- (1) Initialization: Initial settings of the parameter  $\theta$ , which density distribution to be estimated in the hybrid model.  $\mu_k$  is the mean value of each class after clustering using PSO algorithm. Then we can calculate the covariance matrix  $\Sigma_k$ , and  $\alpha_k$ .

- (2) E-step: Calculate the posterior probability  $P_j^n(x_i)$  that each sample point  $i$  belongs to the  $j$  class in the  $n$ th iteration:

$$P_j^n(x_i) = \frac{\alpha_j^n f(x_i | \theta_j^n)}{\sum_{j=1}^k \alpha_j^n f(x_i | \theta_j^n)} \quad (4)$$

- (3) M-step: Get the new parameters when the expected value reaches the maximum by solving the logarithmic likelihood equation. Update the parameter estimates according to:

$$\mu_j^{n+1} = \frac{\sum_{i=1}^n x_i P_j^n(x_i)}{\sum_{i=1}^n P_j^n(x_i)} \quad (5)$$

$$\sum_j^{n+1} = \frac{\sum_{i=1}^n P_j^n(x_i)(x_i - \mu_j^{n+1})(x_i - \mu_j^{n+1})^T}{\sum_{i=1}^n P_j^n(x_i)} \quad (6)$$

$$\alpha_j^{n+1} = \frac{1}{n} \sum_{i=1}^n P_j^n(x_i) \quad (7)$$

- (4) Convergence condition: Repeat E-step and M-step to update the three values above till  $|\theta - \theta'| < \varepsilon$ , where  $\theta'$  is the updated parameters and  $\varepsilon = 10^{-5}$ . Alternatively, the algorithm can be terminated when the velocity updates are close to zero over a number of iterations. Otherwise, switch to step (2).

### 3 Hybrid PSOEM Clustering Algorithm Based on Improved Particle Swarm Algorithm

#### 3.1 The Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a stochastic search process, modeled by the social behavior of a bird flock [13, 14]. In the PSO issue, a swarm refers to a number of potential solutions to the optimization problem, where each particle represents one potential solution. Assuming that each particle occupies a position in  $N$ -dimensional space, it flown through the multi-dimensional search space and adjusted its position to get the particle's best position found thus far and the best position in the neighborhood of that particle. The aim of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function.

Each particle  $i$  concludes the current position of the particle known as  $x_i$ , the current velocity of the particle known as  $v_i$  and the personal best position

of the particle known as  $y_i$ . The particles velocity and position can be adjusted according to:

$$v_{ij}(t+1) = \omega v_{ij}(t) + c_1 r_1(y_{ij}(t) - x_{ij}(t)) + c_2 r_2(\hat{y}(t) - x_{ij}(t)) \quad (8)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (9)$$

where  $\hat{y}_{(t)} \in \{y_0, \dots, y_s\} = \min\{f(y_0(t)), \dots, f(y_s(t))\}$ ,  $s$  is the total number of particle swarm.  $\omega$  is the inertia weight,  $c_1$  and  $c_2$  are the acceleration constants,  $r_1, r_2 \sim U(0, 1)$  and  $j=1,2,\dots,N$ . The velocity is updated based on three parts:

- (1) The fraction of the previous velocity;
- (2) The cognitive component which is a function of the distance of the particle from its personal best position;
- (3) The social component which is a function of the distance of the particle from the best particle found thus far.

The personal best position of particle  $i$  can be updated by:

$$y_i(t+1) = \begin{cases} y_i(t) & \text{if } x_i(t+1) \geq f(y_i(t)) \\ x_i(t+1) & \text{if } x_i(t+1) < f(y_i(t)) \end{cases} \quad (10)$$

Both basic components of the PSO algorithm exist based on the update of the particles' neighborhood. Equation (8) reflects that the neighborhood of global optimum is simply the entire swarm for each particle. The social component enables particles to be drawn toward the best particle in the swarm. In the local optimum part, the swarm is divided into overlapping neighborhoods, and the best particle of each neighborhood is determined. The PSO is usually executed with repeated application of equations (8) and (9) until the specified number of iterations has been exceeded or the velocity updates close to zero over a number of iterations.

### 3.2 The Parameter Optimization of PSO Algorithm

The inertia weight  $\omega$ , a key parameter of PSO algorithm proposed by Shi, can balance the relationship between global searching ability and local searching ability to improve the convergence properties of the algorithm [15]. Large inertia weight is helpful to improve the global searching ability of the algorithm, while a smaller inertia weight will enhance the local searching ability of the algorithm. We expected to find the appropriate inertia weight with optimal balance between global and local searching to reduce the number of iterations while positioning the best solution. Generally speaking, it is hoped that there is a higher pre-search capabilities to get the right seeds in the earlier stage, and a higher development capabilities to speed up the convergence in the later stage.

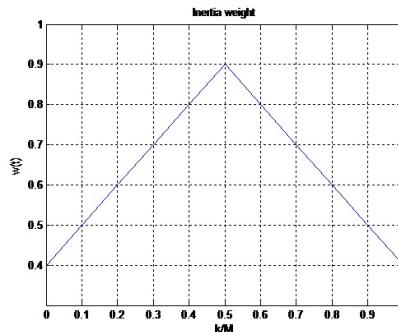
We proposed an improved PSO algorithm in order to get rid of the over-reliance on the initial value and the tendency to get stuck in a local maximum of the standard EM algorithm. More specifically, we cited an inertia weight [16]

increased first and declined after along the fold line according to the relation of parameter constraint between the inertia weight  $\omega$  and the acceleration factor  $c$  under the convergence conditions of the algorithm model to guarantee the constringency of the algorithm. The hybrid PSOEM clustering algorithm based on the improved PSO algorithm and EM clustering algorithm can increase the initialization diversity of EM algorithm and promote optimization search in the global scope. Experimental results of simulation prove its accuracy and validity.

Loan Cristian Trelea and Frans van den Berht pointed out that one requirement needs to be satisfied between the inertia weight  $\omega$  and acceleration factor  $c_1, c_2$  to ensure the algorithms convergence:  $\frac{c_1+c_2}{2} - 1 < \omega$ , i.e.,  $c_1 + c_2 < 2(\omega + 1)$ , or the particles trajectories in the PSO algorithm are divergent. So we accepted the inertia weight increased first and declined after along the fold line:

$$w(t) = \begin{cases} 0.4 + \frac{k}{M} & 0 \leq \frac{k}{M} \leq 0.5 \\ 1.4 - \frac{k}{M} & 0.5 < \frac{k}{M} \leq 1 \end{cases} \quad (11)$$

where  $k$  is the iterations and  $M$  is the maximum iteration. The change of the relationship between them is shown in Fig.1.



**Fig. 1.** Inertia weight varies with the number of iterations

Then combine the acceleration factor with the inertia weight as  $c_1 = (w + 1) * rand_1$ ,  $c_2 = (w + 1) * (2 - rand_2) * rand_2$ , where  $rand_1, rand_2 \sim U(0, 1)$ .

### 3.3 Hybrid PSOEM Clustering Algorithm Based on Improved Particle Swarm Algorithm

EM clustering algorithm is sensitive to the initial value as a local search algorithm, which is easy to converge to a local emxtreum value if the selection of initial value is inappropriate, while PSO algorithm can avoid involving into local extremum for it processing a group of points at the same time rather than being limited to a single point. Therefore, we introduce the PSO algorithm into

EM clustering algorithm as a hybrid PSOEM clustering algorithm based on improved particle swarm algorithm: First of all, we acquire the number of clusters and the initial cluster centers with improved PSO algorithm to get the initial parameters of the mixture model. The mean value of each sample after clustering with PSO algorithm is  $\mu_k$ . Then calculate the covariance matrix  $\sum_k$ , and  $\alpha_k$  is the proportion of each sample in the total samples. The number of clusters is the classification number. We can build a mixed-density model of the image data with multi-iterations of EM algorithm and divide all the pixel values of the image into the corresponding branch of the mixed-density model with Bayesian criterion. The proposed algorithm can increase the initialization diversity of EM algorithm by slowing down the local search and promoting the optimal search over the global scope, so it overcomes the EM algorithms excessive dependence on the initial value and its tendency to get stuck in a local maximum effectively.

For the clustering problem, assume  $N_c$  particles represent  $N_c$  clustering center vector, so each particle  $x_i$  can be presented as:

$$x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{iN}) \quad (12)$$

where  $x_{ij}$  refers to the  $j$ th cluster centroid vector of the  $i$ th particle in cluster  $C_{ij}$ , each cluster acquired by the image clustering is a swarm. The quality of each particle is measured by the following fitness function:

$$f(x_i, Z_i) = v_1 \bar{d}_{\max}(Z, x_i) + v_2 (R_{\max} - d_{\min}(x_i)) \quad (13)$$

In formula (13),  $R_{\max}$  is the maximum pixel value in the image dataset and  $Z$  is the matrix representing the assignment of the patterns to the clusters of the  $i$ th particle. Each element  $z_{ikj}$  indicates whether the pattern  $z_p$  belongs to cluster  $C_{ij}$  of the  $i$ th particle. The constants  $v_1$  and  $v_2$  are user-defined to measure the contributions from each sub-objective. Whereby, the Euclidean distance between the particle and its related clusters is:

$$\bar{d}_{\max}(Z, x_i) = \max_{j \in 1, 2, \dots, N_c} \left\{ \sum_{\forall z_p \in C_{ij}} d(z_p, x_{ij}) / |C_{ij}| \right\} \quad (14)$$

The minimum Euclidean distance between each pair of cluster is:

$$d_{\min}(x_i) = \min_{\forall p, q, p \neq q} \{d(x_{i,p}, x_{i,q})\} \quad (15)$$

where  $|C_{ij}|$  is the number of data vectors that belong to cluster  $C_{ij}$ . The fitness function is a multi-objective optimization problem, which aims to minimize the intra-cluster distance and maximize the inter-cluster separation to reduce the quantization error as much as possible. The priority of the target is different with different initial values of  $v_1$  and  $v_2$ . The algorithm for vector data clustering with the improved PSO algorithm is summarized as follows:

1. Initialize each particle with  $c$  random cluster centers.
2. Calculate Euclidean distance of  $z_p$  with all cluster centroids, assign  $z_p$  to the cluster that have nearest centroid to  $z_p$ . Calculate the fitness function  $f(x_i, Z_i)$  to acquire the personal best and global best position of each particle, update the cluster centroids according to velocity updating and coordinate updating formula.
3. Output the clustering number  $c$ .

The search based on the global space of the PSO algorithm weakens the effect of initial conditions, which is a big improvement of the problem of sensitivity to initial values in K-means clustering. In addition, the multi-thread search raises the convergence speed of PSO algorithm. Specific algorithm process [17] is as follows:

- Step1 Initialize each particle with  $c$  random cluster center vectors;
- Step2 Calculate Euclidean distance of  $x_i$  with all cluster centroids, assign  $x_i$  to the cluster that have nearest centroid to  $x_i$  ;
- Step3 Calculate the fitness function  $f(x_i, Z_i)$  of each particle, and update the globally optimal solution and locally optimal solution;
- Step4 Update the cluster centroids according to the velocity updating and coordinate updating formula;
- Step5 Stop iterating when the terminal condition meets, or turn into step2 and continue;
- Step6 Calculate the overall error  $\varepsilon$ , output the optimal clustering center  $c$  and clustering number when  $\varepsilon < \varepsilon_0$ ;
- Step7 The mean value of each sample after clustering with PSO algorithm is  $\mu_k$ , then calculate the covariance matrix  $\sum_k$ . The classification number is  $k$  and  $\alpha_k$  is the proportion of each sample in the total samples;
- Step8 Conduct the maximum likelihood parameter estimation of each cluster with EM algorithm;
- Step9 Divide all the pixel values of the image into the corresponding branch of the mixed-density model with Bayesian criterion and output the clustering results.

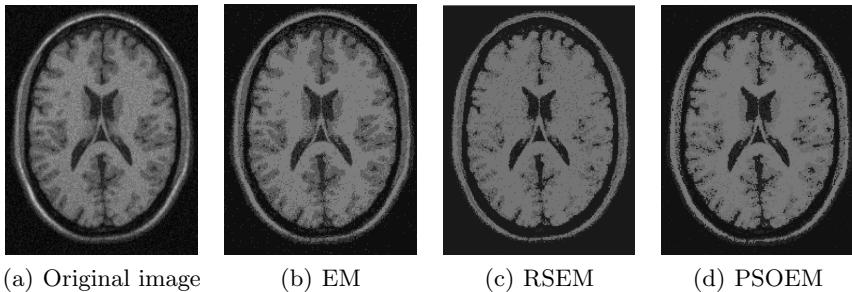
## 4 Experiment and Result Analysis

This section compares the results of the standard EM [18] algorithm and RSEM [8] algorithm with the mixed PSOEM algorithm using the simulated medical images concludes brain image, knee image, and shoulder image to verify the effectiveness of the mixed PSOEM algorithm. The results of clustering are presented in Figure 2 to Figure 4. Figure (a) is the original image, figure(b)(c)(d) are the clustering results obtained by EM algorithm, RSEM algorithm and PSOEM algorithm, respectively.

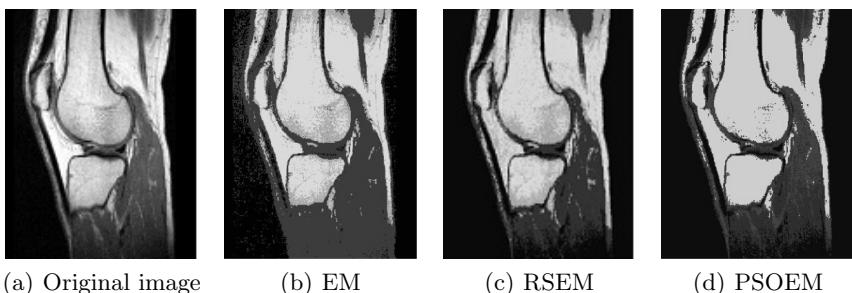
Regards to the parameter setting, the swarm size of PSO algorithm is  $N = 50$ , Each dimension of  $x_{\max}$  and  $x_{\min}$  are the minimum and maximum values of the

sample data respectively, i.e.,  $v_{\max} = |x_{\max}|$ ,  $\varepsilon_0 = 10^{-5}$ , classification number  $c = 4$ , the maximum number of iteration  $M = 200$ . The inertia weight increased first and declined after along the fold line. The involved parameters for each clustering method are set as reasonable value and each experiment repeats 20 times to eliminate the random error.

Subjectively, the experimental results prove that the proposed method obtains clearer boundaries and higher accuracy. The clustering results at the edges in noisy images with poor separability and images with complex internal structure are more accurate. It converges on the target boundary in few iterations.

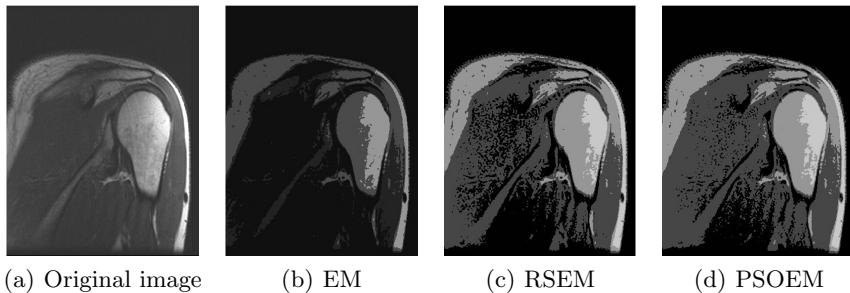


**Fig. 2.** Segmentation results of simulated brain image



**Fig. 3.** Segmentation results of simulated knee image

In order to get the objective evaluation of the proposed algorithm, we adopt F-measure as the evaluation criteria to assess the results, for which combines precision (P) and recall (R). F-measure is a common measure of data selection criteria that widely applied in many information science fields. Clustering accuracy rate (P) indicates the probability that divide the pixels belong to one kind and outside that kind into the same cluster. The higher precision, the more similar pixels concentrated in one cluster. Clustering recall (R) indicates the probability that divide the similar pixels into the same cluster. The higher recall, the more similar pixels concentrated in one cluster indicates the lower probability that they are divided into different clusters.



**Fig. 4.** Segmentation results of simulated shoulder image

Here precision reflects the separating capacity of pixels from different clusters, while recall reflects the recognition capability of pixels from the same cluster. They are caculated by:

$$P = p(i, j) = \frac{C_{ij}}{C_i} \quad (16)$$

$$R = r(i, j) = \frac{C_{ij}}{C_j} \quad (17)$$

where  $C_{ij}$  is the number of particles belongs to the  $i$ th cluster in the  $j$ th cluster.  $C_i, C_j$  represent the number of particles in the  $i$ th and  $j$ th cluster, respectively. F-measure can be acquired by:

$$F = \frac{2PR}{P + R} \quad (18)$$

The larger value of F, the better representation of the boundary.

Figure2 to figure4 are the clustering results with the simulated medical images concludes brain image, knee image, and shoulder image using EM algorithm,

**Table 1.** Comparison of clustering precision and recall

Test Image evaluation index	Clustering method		
	EM	RSEM	PSOEM
Image I	P	0.771	0.750
	R	0.810	0.912
	F	0.790	0.823
Image II	P	0.809	0.827
	R	0.886	0.916
	F	0.846	0.869
Image III	P	0.841	0.848
	R	0.780	0.815
	F	0.809	0.831

RSEM algorithm and PSOEM algorithm, respectively. Table 1 is the comparison of F-measure [19] with three algorithms on three images. As we can see from the table, the clustering accuracy of the proposed algorithm has been greatly improved, and the representation of the boundary is obviously clearer. All in all, the mixed PSOEM algorithm acquires absolute advantage on the clustering quality.

However, the wide improvement of precision will increase the consumption of computation time inevitably. We will focus on improving the arithmetical parameters contain stronger practicality to apply in the actual image processing in the future study.

## 5 Conclusions

We proposed a hybrid clustering algorithm based on the improved particle swarm optimization algorithm and EM clustering algorithm to overcome the shortcomings of EM algorithm, which is sensitive to initial value and easy to sink into local minimum. We apply PSO algorithm to search the global space at the initialization phase to get faster convergence and turn into EM algorithm when the particles in the swarm obtain the global optimal value. The best time for switch is determined by the fitness function. Several trials prove that our proposed method performs better than EM algorithm and RSEM algorithm for its good performance in image clustering. As a future work, we plan to investigate ways to reduce the computational complexity and improve the speed and accuracy of the algorithm.

**Acknowledgements.** The paper was supported by the following fund projects: The natural science foundation of Jiangsu Province(BK20130529); Research Fund for the Doctoral Program of Higher Education of China(20113227110010); Science and technology project of Zhenjiang City(SH20140110); The National Natural Science Foundation of China research on key technology of multimodality medical image processing based on nonparametric density model and rough sets.

## References

- Yao, H., Duan, Q., Li, D., Wang, J.: An improved k-means clustering algorithm for fish image segmentation. *Mathematical and Computer Modelling* 58(3), 790–798 (2013)
- Zhao, F., Fan, J., Liu, H.: Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation. *Expert Systems with Applications* 41(9), 4083–4093 (2014)
- Wu, P., Liu, Y., Li, Y., Shi, Y.: Trus image segmentation with non-parametric kernel density estimation shape prior. *Biomedical Signal Processing and Control* 8(6), 764–771 (2013)

4. Nguyen, T.M., Jonathan Wu, Q., Mukherjee, D., Zhang, H.: A finite mixture model for detail-preserving image segmentation. *Signal Processing* 93(11), 3171–3181 (2013)
5. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631 (2002)
6. Ueda, N., Nakano, R., Ghahramani, Z., Hinton, G.E.: Smem algorithm for mixture models. *Neural Computation* 12(9), 2109–2128 (2000)
7. Verbeek, J.J., Vlassis, N., Kröse, B.: Efficient greedy learning of gaussian mixture models. *Neural Computation* 15(2), 469–485 (2003)
8. Zhao, Q., Hautamäki, V., Kärkkäinen, I., Fränti, P.: Random swap em algorithm for gaussian mixture models. *Pattern Recognition Letters* 33(16), 2120–2126 (2012)
9. Melnykov, V., Melnykov, I.: Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis* 56(6), 1381–1395 (2012)
10. Kennedy, J.: Particle swarm optimization. In: *Encyclopedia of Machine Learning*, pp. 760–766. Springer (2010)
11. Melnykov, V., Maitra, R., et al.: Finite mixture models and model-based clustering. *Statistics Surveys* 4, 80–116 (2010)
12. Liu, Z., Xiao, J.-G., Song, Y.-Q.: Image segmentation based on non-parametric mixture model of legendre orthogonal polynomial. *Jisuanji Yingyong Yanjiu* 27(8), 3165–3167 (2010)
13. Abraham, A., Das, S., Roy, S.: Swarm intelligence algorithms for data clustering. In: *Soft Computing for Knowledge Discovery and Data Mining*, pp. 279–313. Springer (2008)
14. Niasar, N.S., Yazdani, S., Mohajeri, M.: K-nichepso clustering. In: *2008 International Conference on Machine Learning and Cybernetics*, pp. 2668–2672. IEEE (2008)
15. Nickabadi, A., Ebadzadeh, M.M., Safabakhsh, R.: A novel particle swarm optimization algorithm with adaptive inertia weight. *Applied Soft Computing* 11(4), 3658–3670 (2011)
16. Van der Merwe, D., Engelbrecht, A.P.: Data clustering using particle swarm optimization. In: *The 2003 Congress on Evolutionary Computation, CEC 2003*, vol. 1, pp. 215–220. IEEE (2003)
17. Chen, D.-H., Liu, Z.-J., Wang, Z.-H.: Improved possibilistic c-means clustering algorithm based on particle swarm optimization. *Computer Science* 39(11), 122–126 (2012)
18. Yang, M.-S., Lai, C.-Y., Lin, C.-Y.: A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition* 45(11), 3950–3961 (2012)
19. Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al.: A large-scale evaluation of computational protein function prediction. *Nature Methods* 10(3), 221–227 (2013)

# Medical Image Fusion by Combining Nonsubsampled Contourlet Transform and Sparse Representation

Yu Liu<sup>1</sup>, Shuping Liu<sup>1</sup>, and Zengfu Wang<sup>1,2</sup>

<sup>1</sup> Department of Automation, University of Science and Technology of China,  
Hefei 230026, China

<sup>2</sup> Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China  
 [{liuyu1,fengya}@mail.ustc.edu.cn](mailto:{liuyu1,fengya}@mail.ustc.edu.cn),  [zfwang@ustc.edu.cn](mailto:zfwang@ustc.edu.cn)

**Abstract.** In this paper, we present a novel medical image fusion method by taking the complementary advantages of two powerful image representation theories: nonsubsampled contourlet transform (NSCT) and sparse representation (SR). In our fusion algorithm, the NSCT is firstly performed on each of the pre-registered source images to obtain the low-pass and high-pass coefficients. Then, the low-pass bands are merged with a SR-based fusion approach, and the high-pass bands are fused by employing the absolute values of coefficients as activity level measurement. Finally, the fused image is obtained by performing inverse NSCT on the merged coefficients. Several sets of medical source images with different combinations of modalities are used to test the effectiveness of the proposed method. Experimental results demonstrate that our method owns clear advantages over the fusion method based on NSCT or SR individually in terms of both visual quality and objective assessments.

**Keywords:** Medical image fusion, multi-scale transform, nonsubsampled contourlet transform, sparse representation.

## 1 Introduction

Medical images with different imaging modalities reflect different levels of human body information. For instance, the computed tomography (CT) is mostly used in the detection of dense structures like bones and implants, while magnetic resonance imaging (MRI) provides excellent soft-tissue contrast and high-resolution anatomical information. It is practically impossible to capture all the details from one single imaging modality to ensure enough accuracy and reliability of clinical diagnosis. Multimodal medical image fusion [1] offers an important approach to solve this problem by deriving the complementary information from medical images with different modalities. In recent years, multimodal medical image fusion has emerged as an activity research area in medical image analysis and various fusion algorithms have been developed.

Multi-scale transform (MST) theories are the most widely used tools for various image fusion scenarios including medical image fusion. Classical MST-based

fusion methods include pyramid-based ones such as Laplacian pyramid (LP) [2] and gradient pyramid (GP) [3], wavelet-based ones such as discrete wavelet transform (DWT) [4] and dual-tree complex wavelet transform (DTCWT) [5], and multi-scale geometric analysis (MGA)-based ones such as curvelet transform (CVT) [6] and nonsubsampled contourlet transform (NSCT) [7]. In general, the MST-based fusion methods consist of the following three steps [8]. First, decompose the source images into a multi-scale transform domain. Then, merge the transformed coefficients with a given fusion rule. Finally, reconstruct the fused image by performing the corresponding inverse transform over the merged coefficients. These methods assume that the underlying salient information of the source images can be extracted from the decomposed coefficients. Obviously, the selection of transform domain plays a very crucial role in these methods. A comparative study of different MST-based methods is reported in [9], where Li et al. found that the NSCT-based method can generally achieve the best fusion results for multi-modal medical images. In addition to the selection of transform domain, the fusion rule in either high-pass or low-pass band also has a great impact on the fused results. Conventionally, the absolute value of high-pass coefficient is used as the activity level measurement for high-pass fusion. The most popular rule is selecting the coefficient with largest absolute value at each pixel position (the “max-absolute” rule). However, in traditional MST-based fusion methods, the low-pass bands are just simply merged by averaging all the source inputs (the “averaging” rule). Since most energy of an image is contained in the low-pass band, the “averaging” fusion rule tends to cause the loss of contrast in the fused image. In particular, for multimodal medical image fusion, different imaging modalities focus on different components of human body, which may cause that a region in a source image looks very bright while the same region in another source image looks very dark. For example, the bone captured in the CT image owns very high gray level, but the corresponding region in the MRI image owns almost zero-valued gray level. Therefore, when the “averaging” rule is used for low-pass fusion, the contrast of some regions in the fused image will decrease a lot relative to that in the source images.

In the past few years, a new category of image fusion methods based on sparse representation (SR) theory has become a popular topic in image fusion research. SR addresses the signals’ natural sparsity, which is in accord with the physiological characteristics of human visual system [10]. The basic assumption behind SR is that a signal  $\mathbf{x} \in \mathbf{R}^n$  can be approximately represented by a linear combination of a “few” atoms from an overcomplete dictionary  $\mathbf{D} \in \mathbf{R}^{n \times m}$  ( $n < m$ ), where  $n$  is the signal dimension and  $m$  is the dictionary size. That is, the signal  $\mathbf{x}$  can be expressed as  $\mathbf{x} \approx \mathbf{D}\alpha$ , where  $\alpha \in \mathbf{R}^m$  is the unknown sparse coefficient vector. As the dictionary is overcomplete, there are numerous feasible solutions for this underdetermined system. The target of SR is to calculate the sparsest  $\alpha$  which contains the fewest nonzero entries among all feasible solutions (known as sparse coding). Mathematically, the sparest  $\alpha$  can be obtained with the following sparse model.

$$\alpha = \arg \min_{\alpha} \|\alpha\|_0 \quad s.t. \quad \|\mathbf{x} - \mathbf{D}\alpha\|_2 < \varepsilon, \quad (1)$$

where  $\varepsilon > 0$  is an error tolerance and  $\|\cdot\|_0$  denotes the  $l_0 - norm$  which counts the number of nonzero entries.

Yang and Li [11] first introduced SR into image fusion. In their method, the sliding window technique is adopted and the sparse coefficient vector is used as the activity level measurement. In particular, among all the source sparse vectors, the one owning the maximal  $l_1 - norm$  is selected as the fused sparse vector (the “max-L1” rule). The fused image is reconstructed with all the fused sparse vectors. Their experimental results show that the SR-based fusion method own clear advantages over traditional MST-based methods in terms of multifocus image fusion. However, the SR-based fusion method also has its own defects. The fine details in source images like textures and edges tend to be smoothed by the SR-based method for the following two reasons. First, the representation ability of the dictionary may be not sufficient for fine details, which means that the reconstruction result and the input signal is not very similar. As we know, the representation ability of the over-completed dictionary relies much on the number of atoms in it, but a dictionary with a large size will directly increase the computational cost. Moreover, the study in [12] shows that a highly redundant dictionary may lead to potential visual artifacts in the reconstruction result, especially when the input signal is corrupted by noise. Thus, a compromise is usually required. A typical example is that the dictionary size is 256 when the input signal is 64 dimensional ( $8 \times 8$  image patch). Second, the usage of sliding window technique may also cause smoothness. The step size of the sliding window is usually set to 1 when fusing images in spatial domain to avoid undesirable artifacts [11]. However, the more the adjacent patches overlap, the larger the smooth extent. Furthermore, for medical image fusion, there is another disadvantage of the SR-based fusion method. As mentioned before, a same region may be very bright in one source image while very dark in another. In spite of this, the region in both of them may be very flat with few fine details. In this situation, the “max-L1” fusion rule will become very sensitive to the random noise in spatial domain because a small change of value at a pixel may influence the fusion result of several patches. As a result, the fused patches in that region may originate from different source images, which will cause spatial inconsistency in the fused image.

In this paper, we present a new medical image fusion method by taking the complementary advantages of MST and SR. In our algorithm, the NSCT is used as the fusion framework. The low-pass NSCT bands are merged with a SR-based fusion approach, and the high-pass bands are fused using the “max-absolute” rule. The main contribution of this approach is that it overcomes the disadvantages of both NSCT and SR based methods mentioned above. On one hand, with the SR-based fusion approach being used in low-pass bands, more energy in the source images can be preserved, so our method can effectively prevent the loss of contrast. On the other hand, with the high-frequency spatial information being separated by the NSCT, the representation ability of the dictionary is able to meet the reconstruction accuracy, so our method can overcome the inclination of SR-based method to smooth fine details. Moreover, without high-frequency

details, the random noise can be effectively eliminated, so the probability that the patches in a flat region originate from different source images will decrease a lot, leading to better spatial consistency.

## 2 Proposed Fusion Method

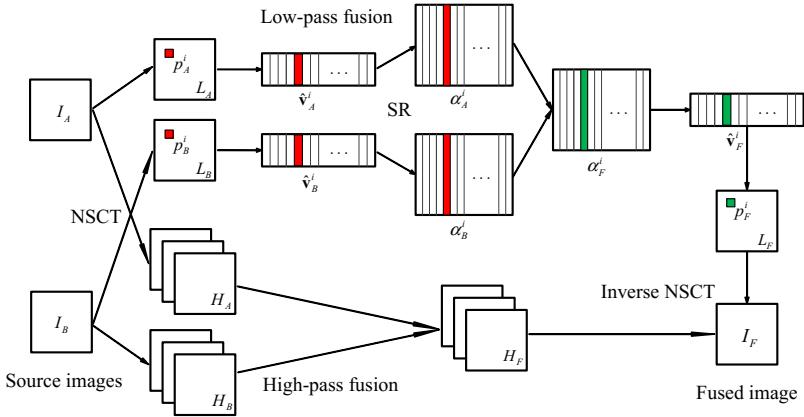
### 2.1 Dictionary Learning

The overcomplete dictionary plays a very important role in the sparse model. Generally, there are two main categories of offline approaches to obtain a dictionary. The first one is directly using the analytical models such as DWT and CWT. However, this category of dictionary is restricted to signals of a certain type and cannot be used for an arbitrary family of signals. The second category is applying the machine learning technique to obtain the dictionary from a large number of training image patches. Suppose that  $M$  training patches of size  $\sqrt{n} \times \sqrt{n}$  are rearranged to column vectors in the  $\mathbf{R}^n$  space, thereby the training database  $\{\mathbf{y}_i\}_{i=1}^M$  is constructed with each  $\mathbf{y}_i \in \mathbf{R}^n$ . The dictionary learning model can be presented as

$$\min_{\mathbf{D}, \{\alpha_i\}_{i=1}^M} \sum_{i=1}^M \|\alpha_i\|_0 \quad s.t. \quad \|\mathbf{y}_i - \mathbf{D}\alpha_i\|_2 < \varepsilon, \quad i \in \{1, \dots, M\}, \quad (2)$$

where  $\varepsilon > 0$  is an error tolerance,  $\{\alpha_i\}_{i=1}^M$  is the unknown sparse vectors corresponding to  $\{\mathbf{y}_i\}_{i=1}^M$  and  $\mathbf{D} \in \mathbf{R}^{n \times m}$  is the unknown dictionary to be learned. The learned dictionaries usually have a better representative ability than the pre-constructed ones, so we used the second approach to obtain a dictionary.

In this work, the sparse model is employed for the fusion of NSCT low-pass bands. One possible way to learn a dictionary is sampling the training patches from NSCT low-pass bands which are decomposed from several images. However, in this situation, the dictionary learning process should be repeated again if either the pyramid filter or the decomposition level of NSCT is changed. Obviously, this will decrease the practicality of the fusion method to a large extent. In this paper, we aim to learn a universal dictionary which can be used in any parameter settings. Since the NSCT low-pass band is obtained by filtering operation on the original image, the low-pass band can be viewed as a smooth version of the original image. Considering that the numerous flat patches contained in a natural image can be well sparsely represented by a dictionary learned from natural image patches, it is feasible to use the same dictionary to represent the patches in the low-pass bands so long as the mean value of each sampled patch is subtracted to zero before training. In this situation, the mean value of each atom in the obtained dictionary is also zero, so the atoms only contain structural information. For an input patch to be represented, its mean value should also be subtracted to zero before sparse coding. Thus, we can directly use natural image patches to learn a universal dictionary.



**Fig. 1.** The schematic diagram of the proposed fusion algorithm

## 2.2 Detailed Fusion Scheme

The schematic diagram of the proposed fusion algorithm is shown in Fig. 1. For simplicity, only the fusion of two source images is considered while the proposed fusion algorithm can be straightforwardly extended to fuse more than two images. The detailed fusion scheme contains the following four steps.

### Step 1: NSCT

Perform NSCT on the two source images  $\{I_A, I_B\}$  to obtain their low-pass bands  $\{L_A, L_B\}$  and high-pass bands which are uniformly denoted as  $\{H_A, H_B\}$ .

### Step 2: Low-Pass Fusion

(i) Apply the sliding window technique to divide  $L_A$  and  $L_B$  into image patches of size  $\sqrt{n} \times \sqrt{n}$  from upper left to lower right with a step length of  $s$  pixels. Suppose that there are  $T$  patches denoted as  $\{p_A^i\}_{i=1}^T$  and  $\{p_B^i\}_{i=1}^T$  in  $L_A$  and  $L_B$ , respectively.

(ii) For each position  $i$ , rearrange  $\{p_A^i, p_B^i\}$  into column vectors  $\{\mathbf{v}_A^i, \mathbf{v}_B^i\}$  and then normalize each vectors mean value to zero to obtain  $\{\hat{\mathbf{v}}_A^i, \hat{\mathbf{v}}_B^i\}$  by

$$\hat{\mathbf{v}}_A^i = \mathbf{v}_A^i - \bar{v}_A^i \cdot \mathbf{1}, \quad (3)$$

$$\hat{\mathbf{v}}_B^i = \mathbf{v}_B^i - \bar{v}_B^i \cdot \mathbf{1}, \quad (4)$$

where  $\mathbf{1}$  denotes an all-one valued  $n \times 1$  vector,  $\bar{v}_A^i$  and  $\bar{v}_B^i$  are the mean values of all the elements in  $\mathbf{v}_A^i$  and  $\mathbf{v}_B^i$ , respectively.

(iii) Calculate the sparse coefficient vectors  $\{\alpha_A^i, \alpha_B^i\}$  of  $\{\hat{\mathbf{v}}_A^i, \hat{\mathbf{v}}_B^i\}$  with the sparse model in Eq. (1) using the OMP algorithm [13].

(iv) Merge  $\alpha_A^i$  and  $\alpha_B^i$  with the “max-L1” rule to obtain the fused sparse vector

$$\alpha_F^i = \begin{cases} \alpha_A^i & \text{if } \|\alpha_A^i\|_1 > \|\alpha_B^i\|_1 \\ \alpha_B^i & \text{otherwise} \end{cases}. \quad (5)$$

The fused result of  $\mathbf{v}_A^i$  and  $\mathbf{v}_B^i$  is calculated by

$$\mathbf{v}_F^i = \mathbf{D}\alpha_F^i + \bar{v}_F^i \cdot \mathbf{1}, \quad (6)$$

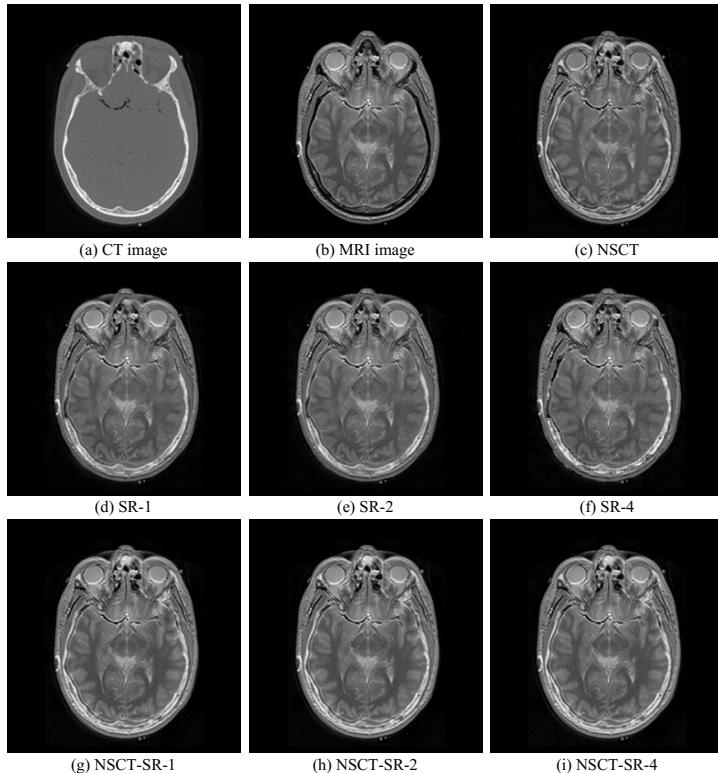
where the merged mean value  $\bar{v}_F^i$  is obtained by

$$\bar{v}_F^i = \begin{cases} \bar{v}_A^i & \text{if } \alpha_F^i = \alpha_A^i \\ \bar{v}_B^i & \text{otherwise} \end{cases}. \quad (7)$$

(v) Iterate the above process for all the source image patches in  $\{p_A^i\}_{i=1}^T$  and  $\{p_B^i\}_{i=1}^T$  to obtain all the fused vectors  $\{\mathbf{v}_F^i\}_{i=1}^T$ . Let  $L_F$  denotes the low-pass fused result. For each  $\mathbf{v}_F^i$ , reshape it into a patch  $p_F^i$  and then plug  $p_F^i$  into its original position in  $L_F$ . As patches are overlapped, each pixel's value in  $L_F$  is averaged over its accumulation times.

### Step 3: High-Pass Fusion

Merge  $H_A$  and  $H_B$  to obtain  $H_F$  with the popular “max-absolute” rule using the absolute value of each coefficient as the activity level measurement. Optionally, apply the consistency verification scheme (see in [4]) to ensure that a fused

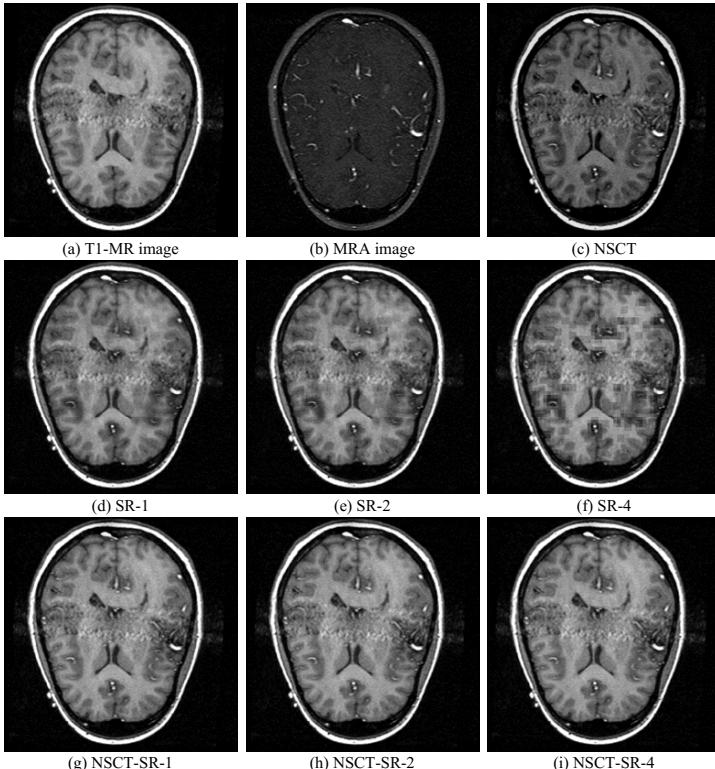


**Fig. 2.** A fusion example of CT and MRI images

coefficient does not originate from a different source image from most of its neighbors. This can be implemented using a small majority filter.

#### Step 4: Inverse NSCT

Perform inverse NSCT over  $L_F$  and  $H_F$  to obtain the final fused image  $I_F$ .



**Fig. 3.** A fusion example of T1-MR and MRA images

## 3 Experiments

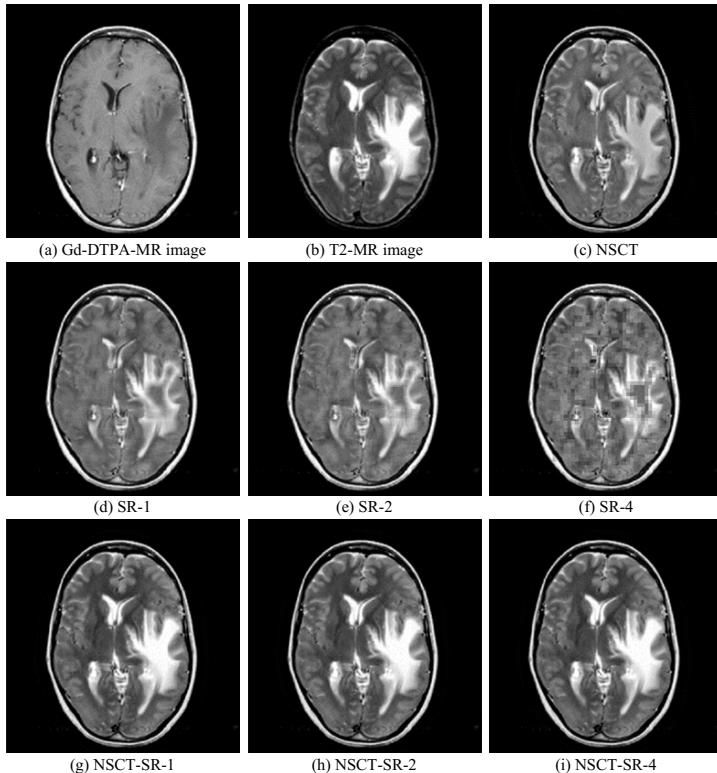
### 3.1 Experimental Setups

To verify the effectiveness of the proposed fusion method (NSCT-SR), the NSCT-based method [7] and SR-based method [11] are mainly used for comparison. Furthermore, for the SR and NSCT-SR methods, the impact of the step length of the sliding window is investigated. In particular, the step lengths are set to 1, 2 and 4, respectively. Thus, there are totally seven methods which are denoted as NSCT, SR-1, SR-2, SR-4, NSCT-SR-1, NSCT-SR-2 and NSCT-SR-4, respectively. The parameters of the above methods are set as follows.

For the NSCT, NSCT-SR-1, NSCT-SR-2 and NSCT-SR-4 methods, we use the ‘pyrexc’ filter as the pyramid filter and the ‘vk’ filter as the directional filter. The decomposition levels are all set to 4, and the direction numbers of the four decomposition levels are selected as 4, 8, 8 and 16, respectively. Moreover, the “max-absolute” rule with a  $3 \times 3$  window based consistency verification scheme [4] is adopted to fuse high-frequency bands for all these four methods. The only difference is the low-frequency bands are merged with the “averaging” rule for the NSCT method. For the SR-1, SR-2, SR-4, NSCT-SR-1, NSCT-SR-2 and NSCT-SR-4 methods, the patch size is set to  $8 \times 8$  and the error tolerance in Eq. (1) is set to 0.1 according to the analysis in [11]. These six methods use a same dictionary which has 256 atoms learned from the K-SVD method [14].

To quantitatively evaluate the performances of different fusion methods, we apply three objective fusion metrics as follows.

1. Standard deviation (*SD*). *SD* measures the overall contrast of the fused image.
2. Entropy (*EN*). *EN* measures the amount of information in the fused image.
3. The gradient based fusion metric  $Q^{AB/F}$  [15].  $Q^{AB/F}$  measures the amount of edge information transferred from the source images to the fused image.



**Fig. 4.** A fusion example of Gd-DTPA-MR and T2-MR images

### 3.2 Experimental Results

Three sets of medical images with different combinations of modalities are tested in our experiments. Fig. 2 shows a fusion example of a CT image and a MRI image. Fig. 3 shows a fusion example of a T1-weighted MR (T1-MR) image and a magnetic resonance angiogram (MRA) image. Fig. 4 shows a fusion example of a MR image after Gd-DTPA (Gd-DTPA-MR) and a T2-weighted MR (T2-MR) image. All the source images have the same size of  $256 \times 256$ .

The fused results of different methods are shown in (c)-(i) of Fig. 2-4. It can be seen that the NSCT method can well preserve the spatial details in the source images, but the overall contrast in its fused images is clearly lower than that in the fused images of both SR-based and NSCT-SR-based methods. On the contrary, the three SR-based methods can obtain fused images with high contrast, but some important spatial details are lost or blurred especially in the fused images of the SR-1 method. However, when the step length becomes larger, which means that the patches are less overlapped, the fused image suffer from more serious blocking effects. In particular, the artifacts cannot be ignored in the fused image of the SR-4 method. Worse still, there exists obvious spatial inconsistency in some flat regions in the SR-based fused images such as in Fig. 4 (d)-(f). The three versions of the proposed NSCT-SR-based method can achieve high-quality results with both high contrast and accurate spatial details. All the important information in the source images can be well injected into the fused image. More importantly, when the step length increases from 1 to 4, the visual quality of the fused image can stably keep high without clear degradation. Thus, compared with traditional SR-based method in which the step length is generally set to 1 or 2, the computational efficiency of the proposed method can be greatly improved by utilizing the NSCT-SR-4 version.

The objective assessments of different fusion methods are listed in Table 1. The maximum in each line shown in bold indicates the best performance over all methods. It can be seen that for all the three fusion examples, the proposed NSCT-SR-based method clearly outperforms the NSCT-based and SR-based methods on all the three metrics. Considering the characteristic of each metric mentioned before, the objective evaluation results is generally in accord with the comparison on visual quality.

**Table 1.** Objective assessments of different image fusion methods

Images	Metrics	NSCT	SR-1	SR-2	SR-4	NSCT-SR-1	NSCT-SR-2	NSCT-SR-4
Fig. 2	$SD$	54.434	55.095	55.268	55.418	58.670	<b>58.671</b>	58.606
	$EN$	5.0999	5.2648	5.2708	5.2562	<b>5.2961</b>	5.2916	5.2916
	$Q^{AB/F}$	0.6134	0.6098	0.6045	0.5943	<b>0.6212</b>	0.6211	0.6209
Fig. 3	$SD$	54.212	65.753	65.863	66.431	69.106	<b>69.121</b>	69.114
	$EN$	5.8618	6.2371	6.2255	5.8496	6.3433	6.3449	<b>6.3494</b>
	$Q^{AB/F}$	0.6144	0.6209	0.6074	0.6098	0.6411	<b>0.6415</b>	<b>0.6415</b>
Fig. 4	$SD$	66.119	67.763	67.892	68.446	69.798	<b>69.829</b>	69.727
	$EN$	4.7586	4.5203	4.5277	4.5570	4.8457	4.8458	<b>4.8496</b>
	$Q^{AB/F}$	0.6322	0.6159	0.6090	0.6035	0.6488	<b>0.6491</b>	0.6486

## 4 Conclusions

In this paper, a novel medical image fusion method by combining NSCT and SR is presented to overcome their respective advantages: the fused images of the NSCT-based method are usually in low contrast while the SR-based method tends to lose spatial details as well as generate spatial inconsistency in the fused image. In our method, the NSCT is employed as the fusion framework. The low-pass NSCT bands are merged with a SR-based fusion approach, and the high-pass bands are fused using the “max-absolute” rule. Experimental results demonstrate that the proposed fusion method can clearly outperform the NSCT- and SR-based methods in terms of both visual quality and objective assessments.

**Acknowledgements.** This work was supported by the National Science and Technology Projects (no. 2012GB102007) and the National Natural Science Foundation of China (No. 61303150).

## References

1. James, A.P., Dasarathy, B.V.: Medical image fusion: A survey of the state of the art. *Information Fusion* 19, 4–19 (2014)
2. Burt, P., Adelson, E.: The laplacian pyramid as a compact image code. *IEEE Transactions on Communications* 31, 532–540 (1983)
3. Petrovic, V.S., Xydeas, C.S.: Gradient-based multiresolution image fusion. *IEEE Transactions on Image Processing* 13, 228–237 (2004)
4. Li, H., Manjunath, B.S., Mitra, S.K.: Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing* 57, 235–245 (1995)
5. Lewis, J.J., OCallaghan, R.J., Nikolov, S.G., et al.: Pixel- and region-based image fusion with complex wavelets. *Information Fusion* 8, 119–130 (2007)
6. Nencini, F., Garzelli, A., Baronti, S., et al.: Remote sensing image fusion using the curvelet transform. *Information Fusion* 8, 143–156 (2007)
7. Zhang, Q., Guo, B.: Multifocus image fusion using the nonsubsampled contourlet transform. *Signal Processing* 89, 1334–1346 (2009)
8. Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion* 4, 259–280 (2003)
9. Li, S., Yang, B., Hu, J.: Performance comparison of different multi-resolution transforms for image fusion. *Information Fusion* 12, 74–84 (2011)
10. Olshausen, B., Field, J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609 (1996)
11. Yang, B., Li, S.: Multifocus image fusion and restoration with sparse representation. *IEEE Transactions on Instrumentation and Measurement* 59, 884–892 (2010)
12. Elad, M., Yavneh, I.: A plurality of sparse representations is better than the sparest one alone. *IEEE Transactions on Information Theory* 55, 4701–4714 (2009)
13. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41, 3397–3415 (1993)
14. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54, 4311–4322 (2006)
15. Xydeas, C.S., Petrovic, V.S.: Objective image fusion performance measure. *Electronics Letters* 36, 308–309 (2000)

# Automated Segmentation and Tracking of SAM Cells

Min Liu and Peng Xiang

College of Electrical and Information Engineering,

Hunan University, Changsha, 410082, China

liu\_min@hnu.edu.cn, peng\_xiang1992@sina.com

**Abstract.** In this paper, we propose an automated segmentation and tracking system for the shoot apical meristem (SAM) cells. Cells are segmented using a mixed filter based watershed segmentation method, which is proved to be very robust and efficient. After segmentation, a Triangle Neighborhood Structure matching method is proposed to track the segmented cells across different time instances. Our tracking method reduces the dependence on neighbors, because we only need two neighbors for any cells for matching while the other local graph matching methods require a much larger number of neighbors. Using our proposed segmentation and tracking system, we are able to track 97% of the plant SAM cells.

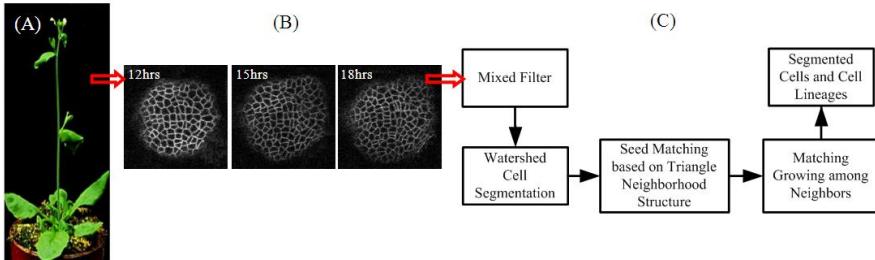
**Keywords:** SAM, Watershed segmentation, Cell tracking, Triangle Neighborhood Structure.

## 1 Introduction

Plant cell research plays an important role in the fields of biology and medicine. Through the analysis of the cell image data such as cell segmentation and tracking, biomedical researchers are able to analyze the cell size distribution and motion characteristics of the cells by tracking the trajectories of cells. Previously, cell segmentation and tracking have been done manually. However, with the increasing of big datasets, manual work is becoming time-consuming and tedious, so automated algorithms for cell segmentation and tracking are attracting great attention in recent years [1,2].

In this paper, we are mainly dealing with plant SAM cells. The SAMs also referred to as the stem-cell niche, is the most important part of the plant body plan because it supplies cells for all the above ground plant parts such as leaves, branches and stem, and at the same time maintains its stable size. The cell images are obtained using Confocal Laser Scanning Microscopy (CLSM) as shown in Figure 1, where (A) is the original plant, and (B) is the cell images taken by CLSM at different time instances. To segment and track the cells in such cell images, the main challenges are the cells' tightly clustered structure and the high noise in the cell images. Lelin Zhang et al. [3] first proposed to build a tracking model based on graph theory, by considering the neighborhood topological relationships among cells to achieve cell identification and tracking. X. Chen et al. [4] proposed to segment and track the cancer cell nuclei using the watershed segmentation. Yanwei Pang et al.[5] proposed a fast and

robust feature extraction algorithm for image matching. Xiaolong Zhou et al.[6] use the GM-PHD filter to tracking multiple moving targets. Those methods are typical segmentation and tracking algorithms, however, they are not dealing well with the tightly clustered SAM cells.



**Fig. 1.** The plant cell images and the proposed cell segmentation and tracking algorithm. (A) The original plant. (B) The cell images collected by the Confocal Laser Scanning Microscopy. (C) The diagram of the proposed segmentation and tracking system.

Given sets of segmented cells at different time instants, cell tracking is essentially a kind of vertex matching problem, which has been widely studied. One of the most popular solutions is the local graph matching method proposed in [7,8]. The watershed method and local graph matching method are able to segment and track most of the cells when the cell images are not highly noised, otherwise, there will be an over-segmentation problem. Moreover, the local graph matching method requires the seed pair to be located in a complete local graph (with all neighboring cells correctly segmented), which cannot be always true, especially in the high noised images.

In order to improve the segmentation accuracy using watershed method, we designed a mixed filter which contains wavelet denoising [9,10] and average filter before final segment. By combining the advantages of spatial filter and frequency filter, most of the noises in the imaging process of SAM cells can be removed; therefore the segmentation accuracy can be highly enhanced. In the tracking procedure, we improved the local graph matching method proposed in [11] by matching a more efficient and compact local structure—the Triangle Neighborhood Structure. The whole segmentation and tracking algorithm is shown in Figure 1(C).

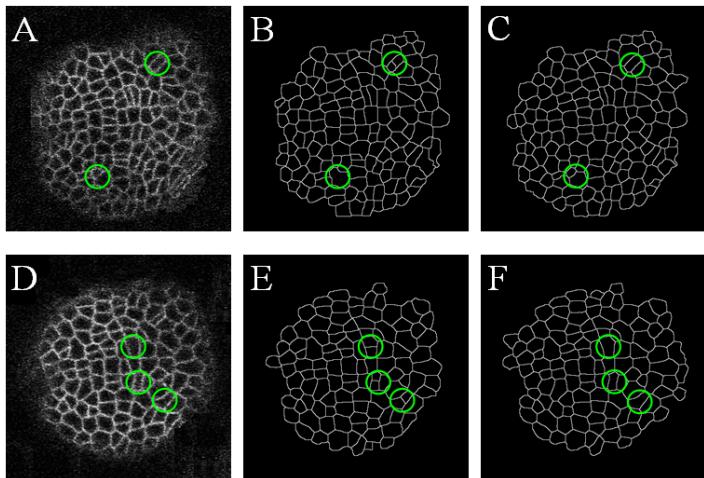
## 2 Detailed Methodology

### 2.1 Watershed Segmentation Based on a Mixed Filter

There are a lot of noises come from the imaging process of SAM cells, as shown in Figure 1 (B). This kind of noise is the main factor of over-segmentation. As pointed out in [11,12], watershed segmentation is a suitable method for SAM cells, but it has an over-segmentation problem. The H-minima method is used to suppress the local minima and it has widely applied to image segmentation [11],[13].The H-minima threshold value plays a crucial role in the watershed algorithm. However, how to choose a suitable threshold to reach the best segmentation result is still a challenge.

Even in the H-minima based segmentation method used in paper [11], the over-segmentation is still a problem. In order to remove such imaging noise, we have designed a mixed filter, which combined the advantages of frequency filter (wavelet denoising filter) and spatial filter (average filter). By using this mixed filter, the over-segmentation in noisy images is overcome effectively, as shown in Figure 2.

Figure 2 shows the difference of segmentation results between using our proposed method and H-minima based method. From this example, we can clearly see that our proposed segmentation method can effectively overcome much of the over-segmentation, as pointed out by the green circles.



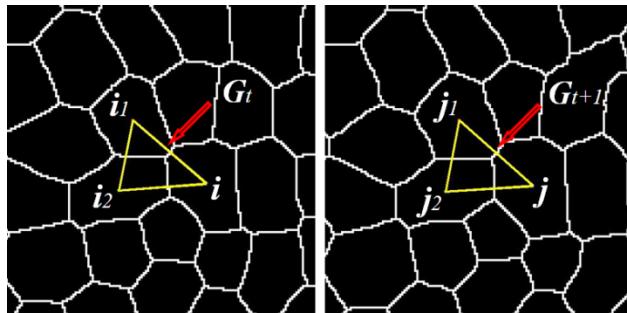
**Fig. 2.** (A) and (D): the original cell images. (B) and (E): the segmentation results using H-minima based watershed segmentation method. (C) and (F): the segmentation results by our proposed method.

## 2.2 The Triangle Neighborhood Structure Matching Method

The local graph matching method[11,12] proposed by Liu Min et al. is widely used to find the cells' correspondences across different cell images. In this method, a seed pair is selected under the condition that two similar local graphs have the same number of neighbor cells. After finding the seed pair, their neighboring cells are gradually matched by a certain local graph matching law, then the newly matched cells act as the new seed pairs and the matching procedure keeps going on till all the possible cells are matched eventually. This method performs well in cell image slightly polluted by noise signal. But for images which are polluted by a lot of noise, we cannot guarantee that the candidate seed pair has the same number of neighbor cells. So in this paper we developed a new method to find the seed pairs based on the matching of the Triangle Neighborhood Structure--TNS, which is the minimum graph in the cell structure. It can improve the tracking accuracy for the above-described tracking procedure and save time.

Before we go to the details of the matching method, we describe here how to create the graphical abstraction given a collection of cells in an image. Every cell is

represented by a vertex in the graph and neighboring vertices are connected by an edge. We regard those cells that are within a certain distance around cell as its neighboring cells, as shown in Figure 3, cells  $i, i_1$  and  $i_2$  are neighbors to each other at time  $t$ , while cells  $j, j_1$  and  $j_2$  are neighbors to each other at time  $t+1$ . Moreover, the cells  $i, i_1, i_2$  or cells  $j, j_1, j_2$  constitute a special local graph structure--triangle neighborhood structure (as  $G_t$  and  $G_{t+1}$  shown in Figure 3), which is the minimum graph representation of cells' neighboring structure.



**Fig. 3.** The Triangle Neighborhood Structures  $G_t$  and  $G_{t+1}$

The local triangle neighborhood structure automatically includes the relative position information of the cells, such as the relative distance between two neighboring cells (the edge length) and the edge orientation. Our proposed tracking algorithm is based on the matching of the triangle neighborhood structures such as  $G_t$  and  $G_{t+1}$  in consecutive cell images.

### Finding the Seed Pair Based on Triangle Neighborhood Structure Matching

For any matching problem, the feature extraction is the most important step in the whole process. In the plant SAM growing process, although the cells are growing and moving outside from the central part, the cells' relative position is not changed much (the images are collected in a reasonable time interval). So the cells' stable position structure is the basic feature used in our tracking algorithm, and the cell matching is based on the matching of the cells' local neighboring structure. In our proposed method, for any triangle neighborhood structure  $G$ , a feature vector  $V$  is extracted to represent this local structure.

Let us consider cell  $i$  as the 'central' cell, which is the rightmost cell in the triangle neighborhood structure  $G_t$  (at time  $t$ ),  $i_1$  and  $i_2$  are the neighborhood cells of  $i$ . There are 3 main components in the feature vector  $V_i(t)$ : the distance between the cells, the relative orientation of the cells and the area of each cell, and this is a 7-length vector as below:

$$V_i(t) = [V_i^1(t), V_i^2(t), V_i^3(t), V_i^4(t), V_i^5(t), V_i^6(t), V_i^7(t)] \quad (1)$$

where  $V_i^1(t)$  and  $V_i^2(t)$  denote the edge length between the center cell  $i$  and the neighboring cells  $i_1$  and  $i_2$  respectively;  $V_i^3(t)$  and  $V_i^4(t)$  are the orientation angles in radians of the edges measured relative to a horizontal axis;  $V_i^5(t)$ ,  $V_i^6(t)$  and  $V_i^7(t)$  denote the area size of three cells  $i$ ,  $i_1$  and  $i_2$  in the triangle neighborhood structure  $G_i$  respectively.

Based on this feature vector description, our proposed seed pair finding method for two consecutive cell images at time  $t$  and at time  $t+1$  is a two-step procedure as below:

#### Step 1: Find the seed pair candidates

For any two triangle neighborhood structures  $G_i$  at time  $t$  and  $G_{i+1}$  at time  $t+1$ . If they are the same triangle structures at time  $t$  and  $t+1$ , their edges' orientations and the cells' area sizes should not change much, which could be mathematically expressed as

$$\begin{cases} |V_i^k(t) - V_j^k(t+1)| \leq T_1, (k = 3, 4) \\ \frac{\max[V_i^w(t), V_j^w(t+1)]}{\min[V_i^w(t), V_j^w(t+1)]} \leq T_2, (w = 5, 6, 7) \end{cases} \quad (2)$$

$T_1$  and  $T_2$  are two thresholds which could be determined in the experiments. The typical values for them in our datasets are 0.175 radians and 1.2. If Equation (2) is satisfied, cells  $(i, j)$  is the candidate seed pair, which will go the step 2 to be verified as a real seed pair or not.

#### Step 2: Find the seed pair based on a distance function

After we find all seed pair candidates from step 1, we should choose the most accurate corresponding cells as the seed pair through a further filtering process based on a distance function of the edges' differences. Given a candidate seed pair  $(i, j)$ , a distance matrix  $D(i, j)$  is established as below

$$D(i, j) = \sqrt{(V_i^1(t) - V_j^1(t+1))^2 + (V_i^2(t) - V_j^2(t+1))^2} \quad (3)$$

In such distance matrix, the element value for any non-candidate pair is assigned a much large value. Let us assume that the first image at time  $t$  has  $N$  cells, and the second image at time  $t+1$  has  $M$  cells. According to the distance function in Equation (3), we can get the distance value  $D(i, j)$  of every cell pair, and pick the most similar cell pair  $(i_s, j_s)$  that satisfies

$$D(i_s, j_s) = \min \{ D(i, j), i = 1 \dots N, j = 1 \dots M \} \quad (4)$$

This seed matching method greatly reduces the dependence on neighbors, because we only need two neighbors for any cells for matching, while the other local graph matching methods [11,12] require an average number of 6 neighbors.

### Tracking Cells from a Seed Pair

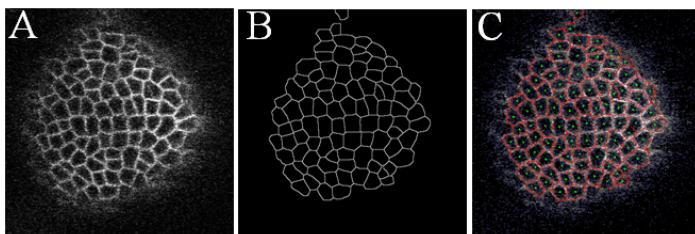
After we find the seed pair, we will continue to match the seed pairs' neighboring cells by a local graph based method which is firstly introduced in [12]. Let  $i$  and  $j$  be two correctly seed cells respectively after matching by the triangle neighborhood structure  $G_t$  at time  $t$  and  $G_{t+1}$  at time  $t+1$ . If  $i_p$  and  $j_q$  are the neighboring cells around them respectively, we have a distance function which is based on the edge's orientation difference and the edge's length difference. The details of this procedure can be found in [12]. This process continues until the algorithm finds all the possible cell matches.

## 3 Experimental Results

We have tested our method on two datasets of SAMs. The experimental results are shown on plant cell images are observed across consecutive time instants, with the time interval of 3 hours between two consecutive instants. Registration is done by the existing method such as the alignment method of Maximization of Mutual Information[14].

### 3.1 Segmentation Results

Figure 4 is an example of the cell segmentation results. A is the original cell image, and B is the segmentation result by our proposed method. In order to verify the segmentation more clearly, we have overlaid the segmentation results on the original images as shown in C, where the red outline represents the segmented cell boundaries, and the green points indicates the centroids of cell regions. From this example, we can see that the segmented cell boundaries are very close to the original cell boundaries.



**Fig. 4.** (A) The original image. (B) The segmentation result by our method. (C) The output image that the cell boundary denoted by red color overlaid on the original image.

We use the accuracy standard introduced in literature [15] to evaluate the segmentation accuracy of our proposed method. In [15],  $F$  is defined as the segmentation accuracy ratio, and  $P$  is the Precision ratio in Equation (6), while  $R$  represents the

Recall ratio in Equation (7).  $N_{GT}$  denotes the image edge pixels by artificial segmentation and  $N_{Det}$  represents the image edge pixels by the algorithm for image segmentation.  $N_{Det \cap GT}$  represents the correct edge pixels detected by algorithm.  $\alpha$  is the weighting factor, which usually takes a value of 0.5.

$$F = \frac{PR}{\alpha R + (1 - \alpha)P} \quad (5)$$

$$P = \frac{N_{Det \cap GT}}{N_{Det}} \quad (6)$$

$$R = \frac{N_{Det \cap GT}}{N_{GT}} \quad (7)$$

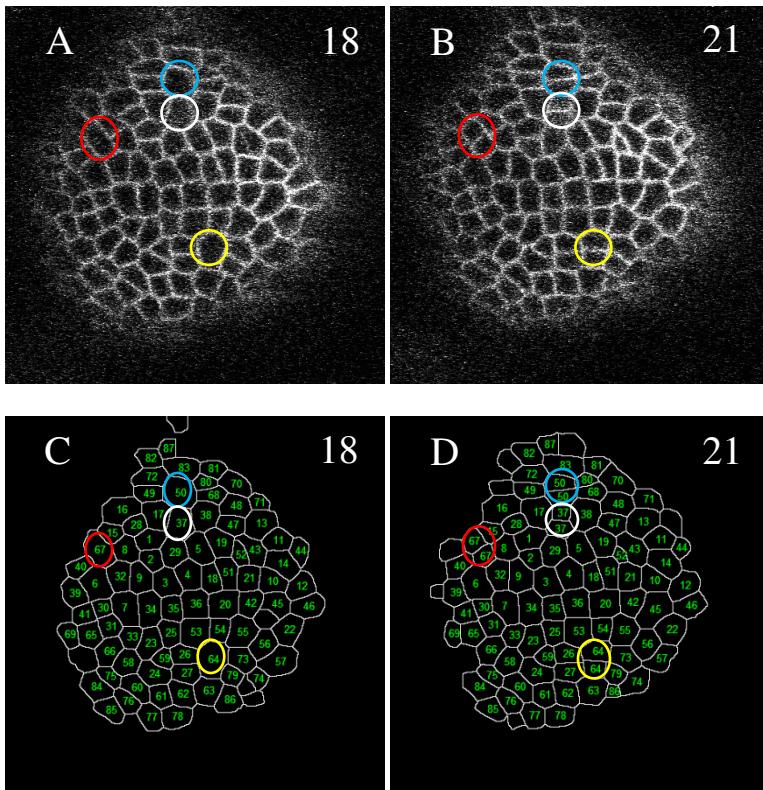
**Table 1.** The number of segmented cells and the segmentation accuracy

<b>Original images</b>	<b>H-minima method</b>		<b>Our method</b>	
	Number of regions	F	Number of regions	F
Image 1	114	0.5697	103	0.6200
Image 2	104	0.6077	92	0.6944
Image 3	112	0.5600	111	0.6104
Image 4	146	0.5169	128	0.5713
Image 5	143	0.5716	133	0.6239

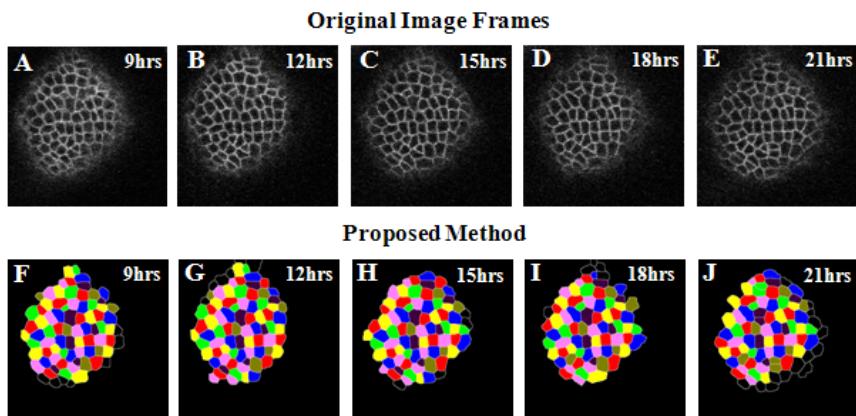
As can be seen from the statistical result in Table 1, the accuracy value  $F$  calculated using the Equation (5) of our proposed method is about 9% larger than the H-minima method. That is to say, the algorithm we proposed has better segmentation accuracy than the method based on H-minima.

### 3.2 Tracking Results

We have tested our proposed matching method on two data sets of SAM. The original images were segmented by our filter based watershed algorithm. Figure 5 is a typical example of the segmentation result and tracking result using our proposed system. The first row is the original images at two time points 18hrs and 21hrs, the second row shows the cells' correspondences using the proposed tracking algorithm. The same cell at those two images is denoted by the same number, from which we could clearly verify the correctness of cells' correspondences.



**Fig. 5.** An example of the proposed segmentation and tracking result. The same number denotes corresponding cells across different time points. A and B are the original image at two time instances: 18<sup>th</sup> and 21<sup>st</sup> hours. C and D are the segmentation and tracking results of A and B using the proposed method.



**Fig. 6.** Segmentation and tracking results for cell images across 5 time instances, where the same cell in different time instances is colored the same

Figure 6 demonstrates the segmentation and tracking results along five time instances (9hrs, 12hrs, 15hrs, 18hrs, and 21hrs). The segmented cells shown in the same color represent the same cell in five consecutive cell images. We have listed the accuracy of tracking method on two datasets, as shown in Table 2, from which we can see that the tracking accuracy is above 97%, which is better than the method proposed in [11], where the tracking accuracy is about 93%.

**Table 2.** The accuracy and the number of cells correctly tracked using the method proposed in [11] and our proposed method

Dataset	Dataset1		Dataset2	
	Method (Liu et al., 2011)	Our method	Method (Liu et al., 2011)	Our method
Number of cells	110	110	115	115
Number of tracked cells	104	108	106	112
Tracking accuracy	94.54%	<b>98.18%</b>	92.17%	<b>97.39%</b>

## 4 Conclusion

In this paper, we present an automated segmentation and tracking system for the plant shoot apical meristem cells. The main challenge comes from the special structure of the SAMs where the cells adhere to each other and high noise from the imaging process. We address the segmentation problem by a local minima filter based watershed segmentation method which proved to be robust to noise, and we track the cells by exploiting the geometric structure and topology of the cells' relative positions. By matching the cells' triangle neighborhood structure we get the cells' correspondences across time. In the experiment, we can segment and track cells with an accuracy of over 97% .

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant 61301254, and the Hunan Provincial Natural Science Foundation of China under Grant 14JJ3069.

## References

1. Barbier de Reuille, P., Bohn-Courseau, I., Godin, C., Traas, J.: A Protocol to Analyse Cellular Dynamics during Plant Development. *The Plant Journal* 44, 1045–1053 (2005)
2. Meijering, E., Dzyubachyk, O., Smal, I.: Methods for Cell and Particle Tracking. *Imaging and Spectroscopic Analysis of Living Cells* 504, 183–200 (2012)
3. Zhang, L., Xiong, H., Zhang, K., Zhou, X.: Graph Theory Application in Cell Nucleus Segmentation, Tracking and Identification. In: *Proceedings of the 7th IEEE International Conference on BioInformatics and BioEngineering*, pp. 226–232 (2007)
4. Chen, X., Zhou, X., Wong, S.T.: Automated Segmentation, Classification, and Tracking of Cancer Cell Nuclei in Time-Lapse Microscopy. *IEEE Transactions on Biomedical Engineering* 53(4), 762–766 (2006)

5. Pang, Y., Li, W., Yuan, Y., Pan, J.: Fully affine invariant SURF for image matching. *Neurocomputing* 85, 6–10 (2012)
6. Zhou, X., Li, Y.F., He, B., Bai, T.: GM-PHD-Based Multi-Target Visual Tracking Using Entropy Distribution and Game Theory. *IEEE Transactions on Industrial Informatics* 10(2), 1064–1076 (2014)
7. Fazl-Ersi, E., Zelek, J.S., Tsotsos, J.: Robust Face Recognition through Local Graph Matching. *Journal Of Multimedia* 2(5), 31–37 (2007)
8. Gold, S., Rangarajan, A.: A Graduated Assignment Algorithm for Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(4), 377–388 (1996)
9. Donoho, D.L., Johnstone, J.M.: Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika* 81(3), 425–455 (1994)
10. Chang, S.G., Yu, B., Vetterli, M.: Spatially Adaptive Wavelet Thresholding with Context Modeling for Image Denoising. *IEEE Transactions on Image Processing* 9(9), 1522–1531 (2000)
11. Liu, M., Chakraborty, A., Singh, D., Yadav, R.K., Meenakshisundaram, G.: Adaptive Cell Segmentation and Tracking for Volumetric Confocal Microscopy Images of a Developing Plant Meristem. *Molecular Plant* 4(5), 922–931 (2011)
12. Liu, M., Roy-Chowdhury, A.K., Gonéhal, V.R.: Exploiting Local Structure for Tracking Plant Cells in Noisy Images. In: *IEEE International Conference on Image Processing*, pp. 1765–1768 (2009)
13. Jung, C., Kim, C.: Segmenting Clustered Nuclei Using H-minima Transform-Based Marker Extraction and Contour Parameterization. *IEEE Transactions on Biomedical Engineering* 57(10), 2600–2604 (2010)
14. Viola, P., Wells III, W.M.: Alignment by Maximization of Mutual Information. In: *Fifth International Conference on Computer Vision*, pp. 16–23 (1995)
15. O’Callaghan, R.J., Bull, D.R.: Combined Morphological-Spectral Unsupervised Image Segmentation. *IEEE Transactions on Image Processing* 14(1), 49–62 (2005)

# Automatic Estimation of Muscle Thickness in Ultrasound Images Based on Revoting Hough Transform (RVHT)

Jianhao Tan<sup>1,\*</sup>, Xiaolong Li<sup>1</sup>, Wentao Zhang<sup>1</sup>, Yaoqin Xie<sup>2</sup>, and Yongjin Zhou<sup>3</sup>

<sup>1</sup> College of Electrical and Information Engineering, Hunan University, China

<sup>2</sup> Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

<sup>3</sup> Shenzhen University, China

lixiaolong19890207@gmail.com

**Abstract.** As an important parameter related to musculoskeletal functions, muscle thickness has been studied for various purposes. However, muscle thickness is usually measured manually by an experienced clinical expert, which is subjective and time consuming, and there are few studies on automatic tracking of muscle thickness during dynamic contraction. In this paper, we proposed a modified Hough transform (HT) to achieve the quantitative and continuous measurement for muscle thickness in ultrasound images. The method involved three steps: image enhancement, locating of superficial and deep aponeuroses by RVHT, and computation of the distance between aponeuroses. The performance of the new method is evaluated using ultrasound images from gastrocnemius muscles of seven patients. The result from the proposed method is also compared to manual detection and another method which was based on Compressive Tracking Algorithm (CTA) applied in our previous work. It was demonstrated in the experiment that the proposed method agrees well with the manual measurement and was able to provide a more convenient and effective approach than the CTA. It could be used for objective muscle thickness tracking in musculoskeletal ultrasound images.

**Keywords:** Muscle thickness, Ultrasound image, Gastrocnemius, Revoting Hough Transform (RVHT).

## 1 Introduction

Ultrasound imaging is effective for imaging soft tissues of the body, such as muscles and tendons [1]. Recently, ultrasound imaging has been employed to measure quantitative muscle changes in morphology [2] [3], such as muscle thickness [4], muscle pennation angle [5], fascicle length [6] and muscle cross-sectional area during contractions [7] [8].

Muscle thickness is probably the most important parameter related to musculoskeletal functions among these morphological parameters [9], and it has been studied in many aspects. For example, Ohata et al. employed muscle thickness to quantify the

---

\* Corresponding author.

muscle strength of people with severe cerebral palsy [10]. English et al. validated that ultrasound was a reliable measure of muscle thickness in acute stroke patients for some anatomical sites [11]. However, muscle thickness was conventionally detected manually in ultrasound images of muscles, and the process is not only subjective, but also time consuming. Subsequently this greatly affects the wider applications of these parameters, particularly for the study of dynamic muscle contraction.

Recently, some researchers have presented several computer-aided methods, which could be used to estimate muscle thickness. Zheng et al. used sonomyography to describe the real-time change of muscle thickness detected using B-mode ultrasound images during its dynamic contraction [12] and proposed to use it for prosthetic control [13]. Koo et al. used cross-correlation to track the locations of aponeuroses and measure muscle thickness on the ultrasound images [14]. However, the cross-correlation method is too sensitive to the size of tracking windows. We also used CTA (more details can be found in [15]) to detect the thickness changes of tibialis anterior muscle during dynamic contraction on ultrasound images, but the initial windows need to be selected manually.

In this paper, we proposed a novel method based on modified Hough transform (HT) [16] to achieve the quantitative and continuous measurement of muscle thickness of gastrocnemius (GM) in ultrasound images. The core concept of the method is as follows: In ultrasound image sequence, we first use a popular method to enhance the hyperechoic regions over the speckles in ultrasonography, namely Multiscale Vessel Enhancement Filtering (MVEF) [17]. Second, the superficial and deep aponeuroses could be located and extracted using RVHT. At last, the muscle thickness was achieved by calculating the distance between the contours of superficial and deep aponeuroses.

## 2 Experiment

Seven healthy male subjects were recruited to participate in this study. No participant had a history of neuromuscular disorders, and all were aware of experimental purposes and procedures. And consent forms were obtained from the subjects prior to the experiment. Human subject ethical approval was granted by the author's institution.

The subjects were in the prone position and performed dorsiflexion/plantar-flexion movements under the direction of examiner. In this process, a real-time B-mode ultrasonic scanner (EUB-8500, Hitachi Medical Corporation, Tokyo, Japan) with a 10-MHz electronic linear array probe (L53L, Hitachi Medical Corporation, Tokyo, Japan) was used to obtain ultrasound images of muscles. The long axis of the ultrasound sound probe was arranged parallel to the long axis of the GM and on its muscle belly. The ultrasound probe was fixed by a custom-designed foam container with fixing straps, and a very generous amount of ultrasound gel was applied to secure acoustic coupling between the probe and skin during muscle contractions, as shown in Fig. 1. The probe was adjusted to optimize the contrast of muscle fascicles in ultrasound images. Then, the B-mode ultrasound images were digitized by a video card (NI PCI-1411, National Instruments, Austin, TX, USA) at a rate of 25 frames/s for later analysis. A total of 5 (subjects)  $\times$  128 (frames) ultrasound images with 323  $\times$  383 pixels were acquired and all the images were cropped to keep the image content only. All data

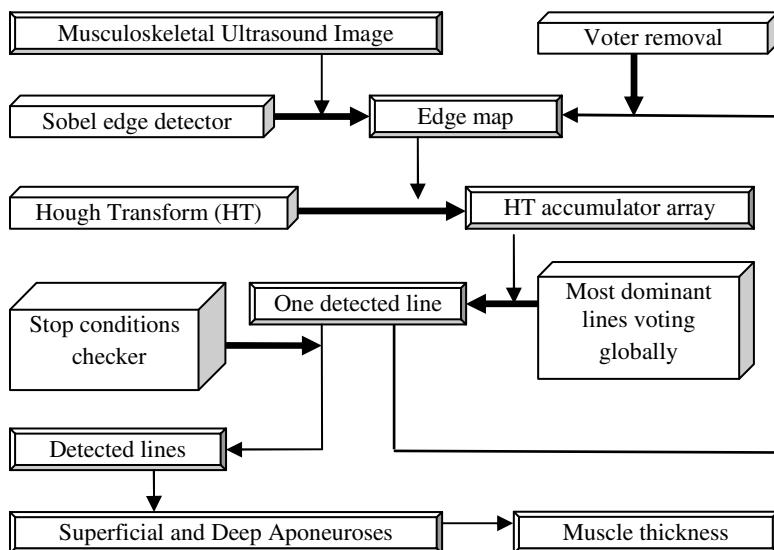
were processed using programs written in matlab (Version R2011b) on a PC equipped with Windows 7, Intel (R) Core T6570 2.10 GHz processors and 2GB RAM.



**Fig. 1.** Experimental setup for collecting ultrasound images from the subject's GM

### 3 Method

The automatic estimation of muscle thickness procedure, based on RVHT method proposed previously, involved three steps: image enhancement using MVEF, locating of superficial and deep aponeuroses by RVHT, and computation of the distance between aponeuroses. Diagram for the procedures is shown in Fig.2.



**Fig. 2.** The diagram of the procedures of the proposed revoting HT

### 3.1 Multiscale Vessel Enhancement Filtering (MVEF)

The MVEF method has good performance in noise and background suppression, which is based on the second order local structure and. The method comprises the following steps: the Hessian matrix estimation (including the choice of Gaussian kernels), computation of eigenvector for each scale and processing for the maximum vesselness response. More details are shown in [17].

### 3.2 Revoting Hough Transform (RVHT)

Standard Hough transform (SHT) uses the normal parameterization of a straight line in an image [18].

$$x \cos\theta_1 + y \sin\theta_1 = \rho_1 \quad (1)$$

Where  $\theta_1$  represents the angle between the x-axis and the normal of the line,  $\rho_1$  is the algebraic distance between the line and the origin. For an arbitrary point  $(x_i, y_i)$  in the image space with coordinates, the lines that go through it are the pairs  $(\rho, \theta)$  with

$$x_i \cos\theta + y_i \sin\theta = \rho \quad (2)$$

These representations correspond to a sinusoidal curve in the  $(\rho, \theta)$  space, which is unique to that point. Then the collinear points will cross each other and an array measuring the crossing situation is accumulated after transforming all edge/feature points to the  $(\rho, \theta)$  space. Traditionally this array  $(\rho, \theta)$  is called accumulator array. The next step of the SHT is an exhaustive search for the maxima in the accumulator array, and all local values of  $H(\rho, \theta)$  exceeding the predefined threshold can be recognized as the evidence of straight lines existing in the original image space.

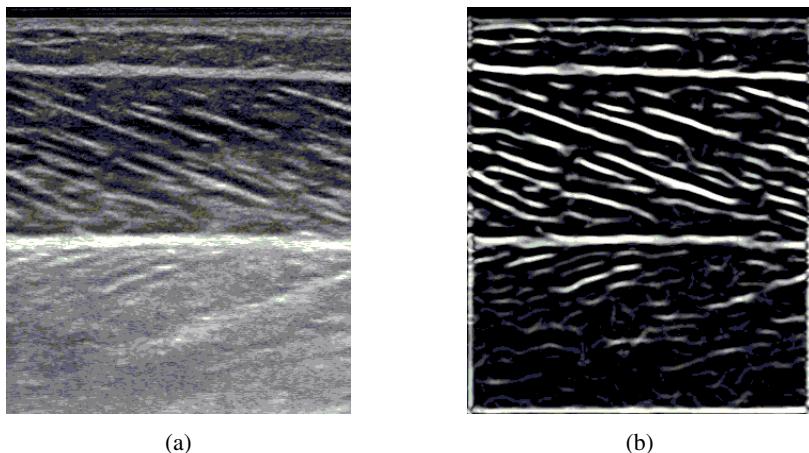
In conclusion, the collinear edge/feature points in the image space show up as peaks in the  $(\rho, \theta)$  space in SHT. However, there arise some issues in the realization of SHT: (1) digital image is by nature discrete; (2) when mapped into  $(\rho, \theta)$  space,  $\theta$  also has to be sampled in a limited resolution and  $\rho$  has to be quantized; (3)  $H(\rho, \theta)$  is also represented on a discrete grid where only integer coordinates have values; (4) The original image can suffer from various noises. These issues could cause problems of aliasing, such as: peak spreading or peak extension.

To test the feasibility of using HT methods for the muscle aponeuroses detection in musculoskeletal sonograms that are usually degraded by speckle noises, a modified HT named as RVHT was adopted in this paper. The RVHT first computes accumulator matrix of Hough transform based on a black-white image called an “edge map”. This edge map represents the meaningful image contents. The RVHT method then locates the global maximum in the accumulator matrix of Hough transform, which corresponds to the most dominant collinear feature points globally, using the standard

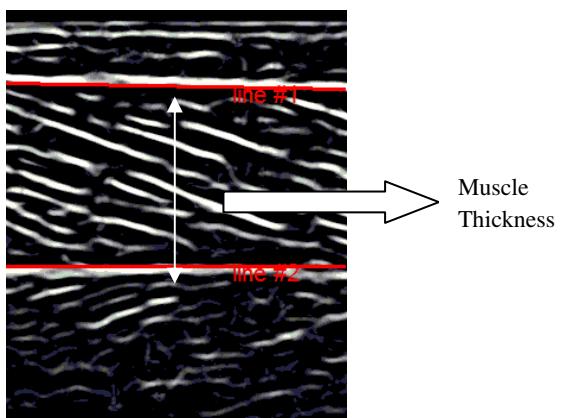
Hough transform. Then the pixels close to the detected line are removed from the edge map and the Hough transform accumulator matrix is calculated again. The same procedure could be executed to search for another line [16].

### 3.3 Estimation of Muscle Thickness

The original images were first cropped to keep the image content only, and the cropped images were then enhanced by MVEF method respectively, as is shown in Fig. 3.



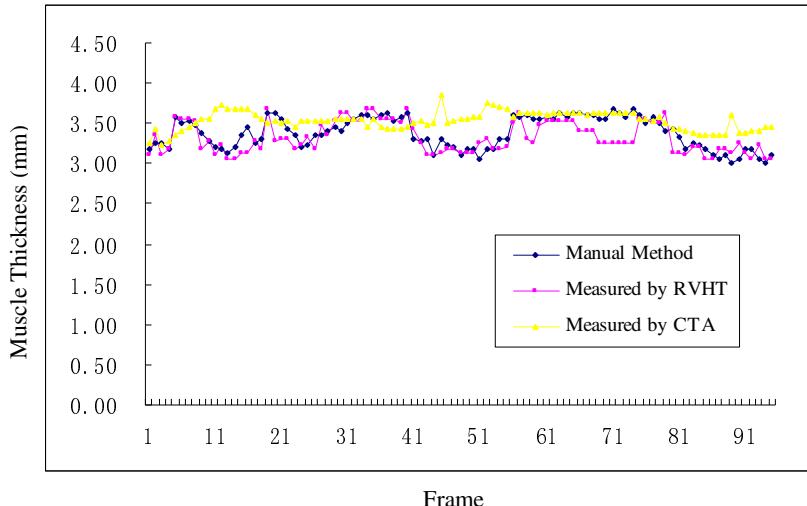
**Fig. 3.** The processing results of gastrocnemius muscle ultrasound images (a) The original image after cropped; (b) the enhancement results of the cropped image after MVEF method



**Fig. 4.** The line detection results using RVHT after MVEF and illustration of muscle thickness definitions. Line #1 and line #2 are recognized as the superficial and deep aponeuroses respectively.

For ultrasound images of skeletal muscles, usually the two of the first few lines detected would be the superficial and deep aponeuroses according to a priori knowledge. And for subject 1 the line #1 and line #2, which is distinctly illustrated in Fig. 4, are recognized as the superficial and deep aponeuroses.

Finally, for each frame, after recognition of the aponeuroses, the mean distance between the superficial and deep aponeuroses is computed as the muscle thickness. As shown in Fig. 4.



**Fig. 5.** Representative comparison result of muscle thickness measured by the Revolving Hough Transform (RVHT), Compressive Tracking Algorithm (CTA) and the manual method (MT)

## 4 Results

The manual measurement is accomplished by one trained clinical expert, the operation is repeated for three times and the results are averaged to obtain the final results. In addition, we also use the CTA method to detect the muscle thickness of the whole image sequence.

For the purpose to make the comparison to the manual method and the CTA, a representative result of muscle thickness of the image sequence, measured by the three methods, is displayed in Fig. 5. It is not hard to see from the curves, this mean muscle thickness, obtained from the results generated using the RVHT, agrees much better with the manual results than the results from CTA.

In order to further evaluate the RVHT method, we define the thickness error rate (TER) to compare muscle thickness measured by the RVHT, CTA and the manual method (MT) as

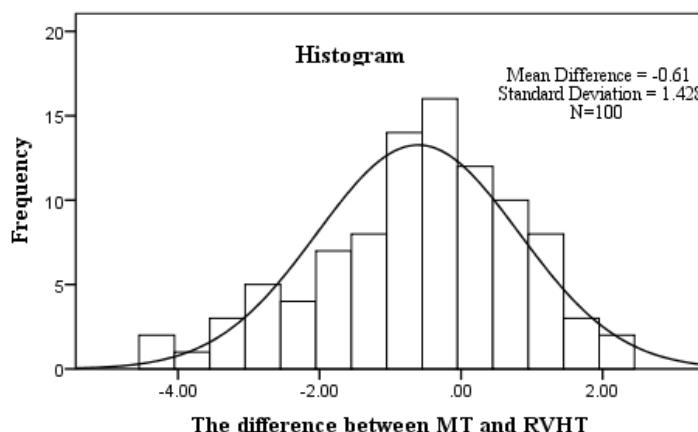
$$TER = |(PT - MT) / MT| \times 100\% \quad (3)$$

where PT is the muscle thickness measured by RVHT or CTA, the average thickness and TER measured for all subjects are listed in Table I.

**Table 1.** Muscle thickness measured by the Revoting Hough Transform (RVHT), Compressive Tracking Algorithm (CTA) and the manual method (MT) for all subjects

Subject	Muscle Thickness (mm)			TER (%)	
	MT	CTA	RVHT	CTA	RVHT
1	33.8±1.9	35.3±1.1	33.2±1.9	3.5±2.8	5.8±5.2
2	36.9±1.4	37.1±1.3	36.4±1.8	4.1±2.8	3.0±2.2
3	37.1±1.6	37.1±1.3	36.8±1.9	2.7±2.1	2.8±2.0
4	45.0±0.7	45.5±0.8	44.3±0.9	2.3±1.9	1.9±1.7
5	37.3±1.3	36.2±0.8	37.1±1.6	3.6±2.5	2.8±2.2
6	37.1±1.1	37.0±1.4	36.6±1.7	4.5±2.8	3.2±2.3
7	36.7±1.3	37.18±1.4	35.6±1.4	4.0±2.9	3.7±2.3

Meanwhile, it can be observed in Table I that the RVHT shows good performance with maximal TER of 5.8% for data from individual subject. The average TER value for all images is 3.26%. Compared to the CTA, the RVHT can provide a pretty good accuracy and efficiency in most cases.

**Fig. 6.** Histogram and the Normal Distribution for the differences between results from the RVHT and manual measurement

Furthermore, SPSS software was used to analysis the difference between results from manual detection and RVHT. It's found that they are not statistically significantly different and the differences fit the normal distribution well, as shown in Fig. 6. It can be concluded that the proposed method is both robust and accuracy, and the accuracy can be

further improved by using better algorithms to enhance the original ultrasound images before RVHT. Certainly, the detecting results could be more precise by improving the quality of acquired images.

## 5 Discussion

As CTA is an established method for estimation of muscle thickness [15], in this work we focus on proposing a new convenient method for automatic estimation of muscle thickness without manually selecting initial tracking windows. It's hoped that this study would provide consultative guide for widespread application of the computerized muscle thickness estimation, thus to replace the time-consuming and subjective manual measurement.

First of all, to automatically detect the muscle thickness, RVHT is used to track lines in ultrasound images of skeletal muscle, and we expect the superficial and deep aponeuroses will be the very first few lines detected. However, without enhancement procedure before RVHT, the performance of aponeuroses detection is quite poor, indicated by the fact that the superficial and deep aponeuroses could not be located as the very first few lines in many images. However, as mentioned before, it becomes much better after image enhancement. In other words, for the automatic estimation of muscle thickness, we must drop the assumption that the superficial and deep aponeuroses are the strongest lines in ultrasound images, unless a proper image enhancement procedure is used.

In future research, more musculoskeletal images should be collected including those under pathological conditions. Problems are expected to arise, and the corresponding further improvements on the automatic estimation of muscle thickness would certainly broaden the application area of ultrasound imaging in clinical musculoskeletal system.

## 6 Conclusions

In this article, we have successfully applied a modified HT using a revoting strategy, aiming at automatic estimation of the muscle thickness in musculoskeletal ultrasound images. The preliminary results obtained with the proposed methods agree well with those obtained using manual measurement and CTA but with much less labor. Results of the experiments suggest that the proposed strategy can be used for objective estimation of muscle thickness in musculoskeletal ultrasound images.

**Acknowledgment.** The work is supported by the next generation communication technology Major project of National S&T (2013ZX03005013), the Low-cost Healthcare Programs of Chinese Academy of Sciences, the Guangdong Innovative Research Team Program (2011S013, GIRTF-LCHT), International Science and Technology Cooperation Program of Guangdong Province (2012B050200004).

## References

1. Hodges, P., Pengel, L., Herbert, R., Gandevia, S.: Measurement of muscle contraction with ultrasound imaging. *Muscle & Nerve* 27(6), 682–692 (2003)
2. Han, P., Chen, Y., Ao, L., Xie, G., Li, H., Wang, L., Zhou, Y.: Automatic thickness estimation for skeletal muscle in ultrasonography: evaluation of two enhancement methods. *BioMedical Engineering OnLine* 12(6) (2013), doi:10.1186/1475-925X-12-6
3. Thoirs, K., English, C.: Ultrasound measures of muscle thickness: intra-examiner reliability and influence of body position. *Clinical Physiology and Functional Imaging* 29(6), 440–446 (2009)
4. Li, J., Zhou, Y.-J., Zheng, Y.-P., Wang, L., Guo, J.-Y.: Sensitive and Efficient Detection of Quadriceps Muscle Thickness Changes in Cross-sectional Plane using Ultrasonography: A Feasibility Investigation. *IEEE Journal of Biomedical and Health Informatics* (2013), doi:10.1109/JBHI.2013.2275002
5. Zhou, Y., Zheng, Y.-P.: Longitudinal enhancement of the hyperechoic regions in ultrasonography of muscles using a gabor filter bank approach: A preparation for semi-automatic muscle fiber orientation estimation. *Ultrasound in Medicine & Biology* 37(4), 665–673 (2011)
6. Loram, I.D., Maganaris, C., Lakie, M.: Use of ultrasound to make noninvasive *in vivo* measurement of continuous changes in human muscle contractile length. *Journal of Applied Physiology* 100(4), 1311–1323 (2006)
7. Muramatsu, T., Muraoka, T., Kawakami, Y., Shibayama, A., Fukunaga, T.: In vivo determination of fascicle curvature in contracting human skeletal muscles. *Journal of Applied Physiology* 92, 129–134 (2002)
8. Guo, J., Zheng, Y.-P., Xie, H., Chen, X.: Continuous monitoring of electromyography (EMG), mechanomyography (MMG), sonomyography (SMG) and torque output during ramp and step isometric contractions. *Medical Engineering & Physics* 32(9), 1032–1042 (2010)
9. Chen, X., Zheng, Y.-P., Guo, J.-Y., Zhu, Z., Chan, S.-C., Zhang, Z.: Sonomographic responses during voluntary isometric ramp contraction of the human rectus femoris muscle. *European Journal of Applied Physiology* 112(7), 2603–2614 (2012)
10. Ohata, K., Tsuboyama, T., Ichihashi, N., Minami, S.: Measurement of muscle thickness as quantitative muscle evaluation for adults with severe cerebral palsy. *Physical Therapy* 86(9), 1231–1239 (2006)
11. English, C., Thoirs, K., Fisher, L., McLennan, H., Bernhardt, J.: Ultrasound is a reliable measure of muscle thickness in acute stroke patients, for some, but not all anatomical sites: A study of the intrarater reliability of muscle thickness measures in acute stroke patients. *Ultrasound in Medicine & Biology* (2012)
12. Zheng, Y.-P., Chan, M., Shi, J., Chen, X., Huang, Q.H.: Sonomography: Monitoring morphological changes of forearm muscles in actions with the feasibility for the control of powered prosthesis. *Med. Eng. Phys.* 28, 405–415 (2006)
13. Guo, J., Zheng, Y.-P., Xie, H., Chen, X.: Continuous monitoring of electromyography (EMG), Mechanomyography (MMG), sonomyography (SMG) torque output during ramp and step isometric contractions. *Medical Engineering & Physics* 32(9), 1032–1042 (2010)
14. Koo, T.K.K., Wong, C., Zheng, Y.: Reliability of sonomyography for pectoralis major thickness measurement. *J. Manipulative Physiol. Therapeutics* 33, 386–394 (2010)
15. Li, X., Li, H., Li, J., Zhou, Y., Tan, J.: Real-Time Estimation of Tibialis Anterior Muscle Thickness from Dysfunctional Lower Limbs Using Sonography. In: Zhang, Y., Yao, G., He, J., Wang, L., Smalheiser, N.R., Yin, X. (eds.) HIS 2014. LNCS, vol. 8423, pp. 63–71. Springer, Heidelberg (2014)

16. Zhou, Y.-J., Zheng, Y.-P.: Estimation of Muscle Fiber Orientation in Ultrasound Images Using Revoting Hough Transform (RVHT). *Ultrasound in Medicine & Biology* 34, 1474–1481 (2008)
17. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
18. Duda, R.O., Peter, E.: The use of Hough transform to detect lines and curves in pictures. *Comm. Assoc. Comp. Hart Machine* 15, 11–15 (1972)

# Influence of Scan Duration on the Reliability of Resting-State fMRI Regional Homogeneity

Xiaotang Li<sup>1</sup>, Jiansong Zhou<sup>2</sup>, and Xiaoyan Liu<sup>1,\*</sup>

<sup>1</sup> College of Electrical and Information Engineering, Hunan University,  
Changsha 410082, China

<sup>2</sup> Mental Health Institute of The Second Xiangya Hospital, Central South University,  
Changsha 410011, China  
xiaoyan.liu@hnu.edu.cn

**Abstract.** Regional homogeneity (ReHo) is widely used in the analysis of fMRI data of patients with schizophrenia. However, the influence of scan duration on the results is not clear. In this work, intraclass correlation coefficient (ICC) was applied to investigate the reliability of a popular method called KCC-ReHo algorithm, using rest-state fMRI data of schizophrenia patients. The full length 6 minutes data collected was split into data with six different durations from 1 min to 6 min with 1 min equal separation. With increasing scan duration, the mean ICC value of the whole brain is found to increase monotonically from 0.55 to 0.97, and the standard deviation decreases from 0.21 to 0.02. The high ICC values mainly occurred in the superior parietal gyrus, paracentral lobule, superior frontal gyrus dorsolateral, supplementary motor area, fusiform gyrus and inferior temporal gyrus of both hemispheres.

**Keywords:** Resting state, Regional homogeneity, ICC, Scan duration, Schizophrenia.

## 1 Introduction

Recently, there is an increasing number of studies investigating the neural activity by means of resting-state functional magnetic resonance imaging (Rs-fMRI) since the first Rs-fMRI study by Biswal[1, 2]. From then on, Rs-fMRI has been widely performed in patients with some psychiatric disorders, such as schizophrenia[3-5].

One of the popular methods used to analyze the data is so called KCC-ReHo algorithm[6-10], which calculates the regional homogeneity (ReHo) of resting state time series based on the Kendall's coefficient of concordance (KCC) among one voxel and its 26 neighbors [11]. KCC-ReHo method is a non-parametric data-driven approach and robust against noise [12], and its effectiveness was also demonstrated in our previous work[13]. However, it is also reported that the scan duration of fMRI could affect the analysis results [14, 15]. A recent study by Zuo[16] showed that the KCC-ReHo method would benefit from longer duration acquisitions. The objective of the

---

\* Corresponding author.

present work is to investigate the influence of the scan duration on the reliability of a popular method called KCC-ReHo algorithm, using rest-state fMRI data of schizophrenia patients.

## 2 Materials and Methods

### 2.1 Scan Acquisition

The resting-state fMRI data of twenty-six male schizophrenia patients were collected on a SIEMENS AVANPO 1.5T scanner. A birdcage head coil was used to minimize head motion. Participants were asked to relax and remain still with their eyes closed, refrain from any cognitive, language, or motion as much as possible, and not to fall asleep. All functional images had 3.75 mm isotropic voxels and were acquired with an echo planar imaging (EPI) sequence: TR = 3000 ms, TE = 50 ms, flip angle = 90°, FOV = 240 mm × 240 mm, matrix size = 64 × 64, slice thickness = 3 mm, and number of slices = 29. T1-weighted structural images were collected prior to functional scans with an SE sequence: TR = 500 ms, TE = 11 ms, flip angle = 70°, FOV = 240 mm × 240 mm, matrix size = 144 × 192, slice thickness = 3 mm, and number of slices = 29. For each subject, the Rs-fMRI scanning lasted for 6 min and 120 volumes were obtained.

### 2.2 Data Preprocessing

Data preprocessing was performed with MATLAB using the statistical parametric mapping software toolkit (SPM8). The collected 120 volumes were corrected to reduce the influence of head movement. All subjects should have no more than 1.5 mm maximum displacement in x, y, or z and 1.5° of angular motion. Afterwards, the functional images were normalized and resampled to 3 × 3 × 3 mm<sup>3</sup>. Then the images were linearly detrended and temporally bandpass filtered (0.01–0.08 Hz) to reduce the effect of low-frequency drifts and physiological high-frequency noise. Details about the subjects and data acquisition procedures can be found in [17] and will be therefore not repeated here.

### 2.3 KCC-ReHo Algorithm

The regional homogeneity (ReHo) method is one of the most popular methods for detecting the similarity or synchronous of the time series of a given voxel to those of its nearest neighbors in the Rs-fMRI. For a given voxel, KCC-ReHo is defined as:

$$W = \frac{\sum_{i=1}^n (R_i)^2 - n(\bar{R})^2}{\frac{1}{12}K^2(n^3 - n)} \quad (1)$$

where  $K$  is the number of neighbors of the voxel (including the voxel, a total of 27 voxels was used in this study),  $n$  is the number of volumes,  $R_i$  is the rank across its neighbors at the  $i$ -th time point, and  $\bar{R}$  is the overall mean rank across all neighboring voxels and volumes.

All individual voxel-wise KCC-ReHo values were computed and standardized into KCC-ReHo m-values by dividing by the mean voxel-wise KCC-ReHo obtained for the entire brain for subsequent analyses. The dimensionless parameter is introduced:

$$mReHo = \frac{W}{\bar{W}} \quad (2)$$

where  $W$  is the KCC-ReHo values and  $\bar{W}$  is the mean KCC value of the entire brain.

With above definitions, an individual mReHo map for each subject was obtained, which was then processed by Gaussian spatially smoothing (4mm full width at half maximum (FWHM)) in order to suppress noise, and the mReHo map became the smReHo map.

## 2.4 Reliability of the KCC-ReHo

In this work, intraclass correlation coefficient (ICC) was calculated to investigate the reliability of KCC-ReHo algorithm mentioned above. The full length 6 min data was split into data with six different durations from 1 min to 6 min with 1 min equal separation. Then, we examine the reliability and similarity of KCC-ReHo smoothed (smReHo) for scans ranging in length from 1 to 5 min versus 6 min. For each brain voxel, the KCC-ReHo was first rearranged into  $26 \times 2$  matrices. Here, the  $26 \times 2$  matrices represented smReHo of the 26 subjects across scans with 2 different lengths of volumes in a single run (different scan duration reliability). Using a one-way ANOVA on each of the two matrices, with random subject effects, we split the mean of the squares into between-subject mean squares (MS<sub>b</sub>) and within-subject mean squares (MS<sub>w</sub>). ICC values were subsequently calculated according to the following equation where  $k$  is the number of repeated observations per subject [18]:

$$ICC = \frac{MS_b - MS_w}{MS_b + (k - 1)MS_w} \quad (3)$$

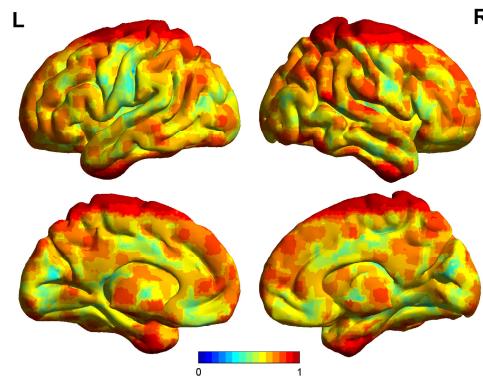
As per equation (3), for a measure to be reliable (exhibiting high ICC) there should be low within-subject variance relative to between-subject variance. Thus, ICC ranges from 0 (no reliability) to 1 (perfect reliability). The ICC values were categorized into five common intervals <sup>[19]</sup>:  $0 < ICC \leq 0.2$  (slight),  $0.2 < ICC \leq 0.4$  (fair),  $0.4 < ICC \leq 0.6$  (moderate),  $0.6 < ICC \leq 0.8$  (substantial), and  $0.8 < ICC < 1.0$  (almost perfect).

### 3 Results and Discussions

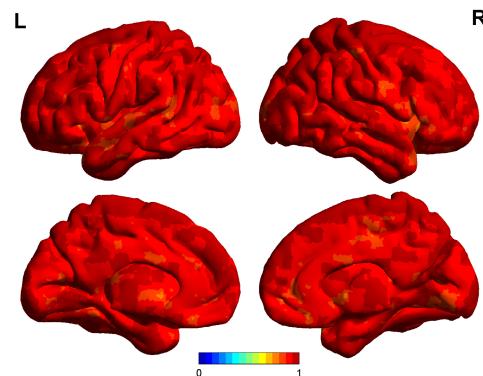
Table 1 gives the calculated ICC values for the five cases of 1versus 6, 2versus 6 3 verus 6, 4versus 6, 5versus 6. The mean ICC value of the whole brain shows to increase monotonically from 0.55 to 0.97 and the standard deviation (std) decrease monotonically from 0.21 to 0.02, as the scan duration increases from 1 minutes to 5 minutes. It seems that the reliability of KCC-ReHo algorithm increases significantly with longer scan durations.

**Table 1.** The ICC among Different Scan Durations

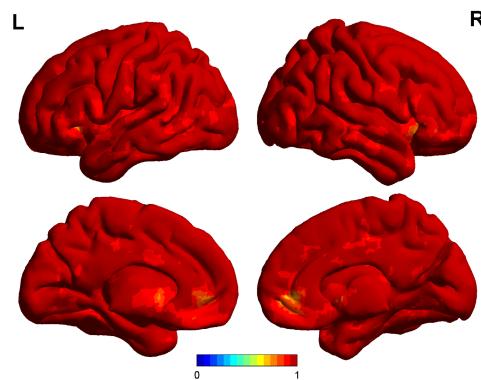
Group(min)	1vs6	2vs6	3vs6	4vs6	5vs6
Mean-ICC	0.55	0.75	0.85	0.92	0.97
Std-ICC	0.21	0.14	0.09	0.05	0.02



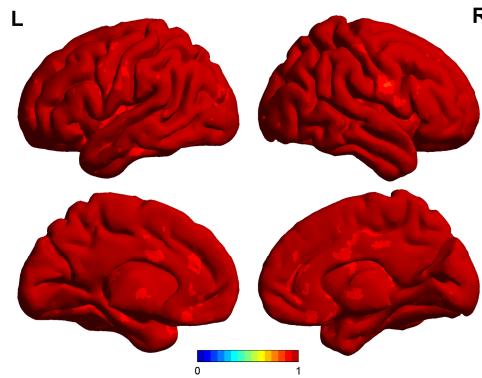
**Fig. 1.** ICC map for the case of 1min vs 6 min



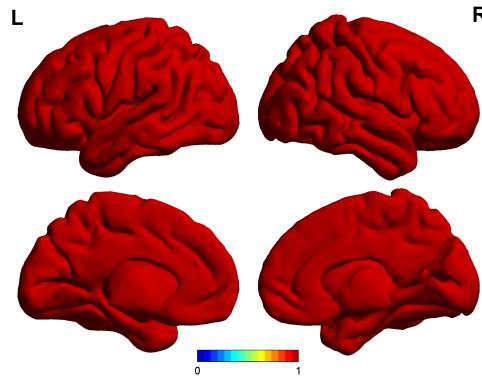
**Fig. 2.** ICC map for the case of 2min vs 6 min



**Fig. 3.** ICC map for the case of 3min vs 6 min



**Fig. 4.** ICC map for the case of 4min vs 6 min



**Fig. 5.** ICC map for the case of 5min vs 6 min

The ICC maps corresponding to the above five cases are shown in Fig. 1- Fig.5. It can be seen that the persistent high ICC value mainly occurred in the superior parietal gyrus, paracentral lobule, superior frontal gyrus dorsolateral, supplementary motor area, fusiform gyrus and inferior temporal gyrus of both hemispheres.

## 4 Conclusions

In this work, ICC was used to investigate the effect of scan duration on the reliability of KCC-ReHo algorithm for rest-stating fMRI data analysis. It is shown that the reliability of KCC-ReHo increases significantly with longer scan durations. In our study, a high ICC value of 0.92 was reached after four minutes. In our future work, more Rs-fMRI data will be collected to verify this finding.

**Acknowledgements.** Financial support from NSFC China (61374149), Hunan Provincial Natural Science Foundation of China (13JJA003), and Research Fund for the Doctoral Program of Higher Education (20130161110010) were greatly appreciated.

## References

1. Biswal, B., Zerrin Yetkin, F., et al.: Functional connectivity in the motor cortex of resting human brain using echo - planar mri. *Magnetic Resonance in Medicine* 34(4), 537–541 (1995)
2. Lang, S., Duncan, N., et al.: Resting-State Functional Magnetic Resonance Imaging: Review of Neurosurgical Applications. *Neurosurgery* 74(5), 453–465 (2014)
3. Karbasforoushan, H., Woodward, N.D.: Resting-State Networks in Schizophrenia. *Current Topics In Medicinal Chemistry* 12(21), 2404–2414 (2012)
4. Liu, H., Kaneko, Y., et al.: Schizophrenic patients and their unaffected siblings share increased resting-state connectivity in the task-negative network but not its anticorrelated task-positive network. *Schizophr Bull.* 38(2), 285–294 (2012)
5. Yu, Q., Allen, E.A., et al.: Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current Topics In Medicinal Chemistry* 12(21), 2415–2425 (2012)
6. Lopez-Larson, M.P., Anderson, J.S., et al.: Local brain connectivity and associations with gender and age. *Developmental Cognitive Neuroscience* 1(2), 187–197 (2011)
7. Dai, X.J., Gong, H.H., et al.: Gender differences in brain regional homogeneity of healthy subjects after normal sleep and after sleep deprivation: a resting-state fMRI study. *Sleep Medicine* 13(6), 720–727 (2012)
8. Wang, L., Song, M., et al.: Regional homogeneity of the resting-state brain activity correlates with individual intelligence. *Neuroscience Letters* 488(3), 275–278 (2011)
9. Zuo, X.N., Di Martino, A., et al.: The oscillating brain: complex and reliable. *Neuroimage* 49(2), 1432–1445 (2010)
10. Yan, C.G., Zang, Y.: DPARSF: a MATLAB toolbox for pipeline data analysis of resting-state fMRI. *Frontiers in Systems Neuroscience* 4 (2010)
11. Zang, Y., Jiang, T., et al.: Regional homogeneity approach to fMRI data analysis. *NeuroImage* 22(1), 394–400 (2004)

12. Zuo, X.N., Xing, X.: Test-retest reliabilities of resting-state fMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews* (2014)
13. Zhu, P., Liu, X.Y., et al.: fMRI analysis of schizophrenic patients with aggressive behavior. In: Chinese Automation Congress(CAC), 675–678. IEEE (2013)
14. Tian, L., Ren, J., et al.: Regional homogeneity of resting state fMRI signals predicts Stop signal task performance. *NeuroImage* 60(1), 539–544 (2012)
15. Birn, R.M., Molloy, E.K., et al.: The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage* 83, 550–558 (2013)
16. Zuo, X.-N., Xu, T., et al.: Toward reliable characterization of functional homogeneity in the human brain: Preprocessing, scan duration, imaging resolution and computational space. *NeuroImage* 65, 374–386 (2013)
17. Yi, J.L., Wang, X.P., et al.: Magnetic resonance imaging study of smygdala functional connectivity pattern of male schizophrenic patients with aggressive behavior. *Chinese Journal of Clinical Psychology* 17(6), 669–671 (2009)
18. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86(2), 420–428 (1979)
19. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159–174 (1977)

# A Global Eigenvalue-Driven Balanced Deconvolution Approach for Network Direct-Coupling Analysis

Haiping Sun and Hongbin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University,  
and Key Laboratory of System Control and Information Processing,  
Ministry of Education of China, Shanghai, 200240, China  
[hbshen@sjtu.edu.cn](mailto:hbshen@sjtu.edu.cn)

**Abstract.** It is an important and unsettled issue to distinguish direct dependences from the indirect ones without any prior knowledge in biological networks and social networks, which contain important biological features and co-authorship information. We present a new algorithm, called balanced network deconvolution (BND), by exploiting eigen-decomposition and the statistical behavior of the eigenvalues of random symmetric matrices. Specially, the BND is a parameter-free algorithm that can be directly applied to different networks. Experimental results establish BND as a robust and general approach for filtering the transitive noise on various input matrices generated by different prediction algorithms.

**Keywords:** Network direct-coupling, BND, Eigenvalue transformation, Transitive noise model.

## 1 Introduction

Many complex systems, such as social [1], biological [2] and information sciences [3], can be abstracted as networks that only include nodes and edges. Although much progress has been obtained on complex network studies, the complexity of modern real-world network data is far complicated beyond our initial thoughts. For instance, one of the challenges is the observed network is noisy and inaccurate due to the data is contaminated by the variable indirect relationship [4, 5] due to the limited accuracy of the current methods. Since the network data quality will significantly affect the reliability of results in the following steps, separating the direct from indirect contacts is thus an essential but tough task, especially in the shortage of prior knowledge.

The transitive effects of correlations are considered as a main source of indirect contacts [4]. If there are true contacts between sites AB and BC, it will result in false observed correlations between AC with a high probability. This kind of noise exists widely in different kinds of networks including protein residue contact network, gene regulatory network and social network. Many groups have developed different approaches to find direct information flows, e.g. graphical models [6], Bayesian networks [7], and the message-passing algorithms [8]. In protein residue contact network area, DI[9] and MI[10] are two popular methods used to derive true residue contact.

And CLR[11], ARACNE, MI[10], Pearson, Spearman[12], GENIE3[13], TIGRESS[14], Interelator[15], ANOVerence[16] and Community are proposed to reconstruct the gene regulatory network. But these methods have limited applicability because of their specific design for target fields. Compared to them, we propose a general and parameter-free algorithm to distinguish direct dependencies from the indirect ones cross different kinds of network.

## 2 Related Work and Drawbacks

Network deconvolution (ND) [4] is cutting-edge method on removing the indirect correlations with general applicability. It formulates the transitive closure of a network as an infinite sum of true direct network, which can be written in a closed infinite-series sum. The relationship between the true network matrix ( $\mathbf{G}_{\text{dir}}$ ) and the observed network matrix ( $\mathbf{G}_{\text{obs}}$ ) in ND algorithm is shown in Eq.1. After matrix eigenvalue decomposition, the relationship between  $\lambda_{\text{obs}}$  and  $\lambda_{\text{dir}}$  can be derived as shown in Eq.2.

$$\mathbf{G}_{\text{obs}} = \mathbf{G}_{\text{dir}} + \mathbf{G}_{\text{dir}}^2 + \mathbf{G}_{\text{dir}}^3 + \dots \quad (1)$$

$$\lambda_{\text{obs}} = \frac{\lambda_{\text{dir}}}{1 - \lambda_{\text{dir}}} \quad (2)$$

where  $\lambda_{\text{obs}}$  is the eigenvalue of the observed network, and  $\lambda_{\text{dir}}$  the eigenvalue of the true network.

Actually, to sum the infinite series of  $\mathbf{G}_{\text{dir}}$  (Eq.1),  $\lambda_{\text{dir}}$  must satisfy the condition that  $\max(|\lambda_{\text{dir}}|) < 1$ . However, when under this condition, according to Eq.2, all the eigenvalues of the  $\mathbf{G}_{\text{obs}}$  must be  $\lambda_{\text{obs}} > -0.5$ . So there needs a tuning parameter to linearly scale  $\lambda_{\text{obs}}$ , observed from real-world applications, (Eq.3, Eq.4) in the original ND model.

$$\lambda_{\text{dir}} = \frac{\alpha \lambda_{\text{obs}}}{1 + \alpha \lambda_{\text{obs}}} \quad (3)$$

$$\alpha \leq \min\left(\frac{\beta}{(1-\beta)\lambda_{\text{obs}}^{+(\text{max})}}, \frac{-\beta}{(1+\beta)\lambda_{\text{obs}}^{-(\text{min})}}\right) \quad (4)$$

where  $\lambda_{\text{obs}}^{+(\text{max})}$  means the biggest positive  $\lambda_{\text{obs}}$  eigen-value and  $\lambda_{\text{obs}}^{-(\text{min})}$  means the smallest negative  $\lambda_{\text{obs}}$ .

The  $\beta$  parameter, which is network dependent, restrains the largest absolute value of  $\lambda_{\text{dir}}$ . When applied to residue contact network, the bigger  $\beta$  is, the better performance is. So a recommended  $\beta$  values was 0.99. 0.5 was suggested on the gene regulator network instead. Because the performance line on this database is convex and ND achieves best performance when  $\beta$  is 0.5. But  $\beta$  doesn't influence the performance on co-authorship network. The author used 0.95 on this network.

### 3 Proposed Method

To solve the parameter-dependent problem, in this paper, we proposed a new balanced network deconvolution (BND) algorithm to remove transitive relationships. The core part of ND is its noise model and the transformation of eigenvalues. So we inspect its noise model (Eq.1) carefully, finding that it contains the even and odd powers of direct matrix. The odd powers will keep the plus or minus signs of the eigenvalues after transformation, while the even powers will arbitrarily make all transformed eigenvalues positive, consequently making the eigenvalue distributions imbalanced. Besides, according to Wigner's semi-circle law[17], the distribution of eigenvalues from symmetric random matrices, whose entries obey normal Gaussian distribution is balanced distributed like a semi-circle. So instead of representing the observed matrix data as the sum of true matrix and all its powers, we only consider the true matrix and its odd powers as the noise model. The noise model of BND is as follows:

$$\mathbf{G}_{\text{obs}} = \mathbf{G}_{\text{dir}} + \mathbf{G}_{\text{dir}}^3 + \mathbf{G}_{\text{dir}}^5 + \dots \quad (5)$$

By summing the infinite series in Eq.5, we get the closed form as Eq.6.

$$\mathbf{G}_{\text{obs}} = \mathbf{G}_{\text{dir}} (\mathbf{I} - \mathbf{G}_{\text{dir}}^2)^{-1} \quad (6)$$

Then we use  $\mathbf{U}$  and  $\mathbf{E}_{\text{dir}}$  to represent eigenvectors and a diagonal matrix of eigenvalues from  $\mathbf{G}_{\text{dir}}$ , where  $\lambda_{\text{dir}}^i$  is the  $i$ -th diagonal component of the matrix  $\mathbf{E}_{\text{dir}}$ . By using the eigen decomposition principle, we have  $\mathbf{G}_{\text{dir}} = \mathbf{U}\mathbf{E}_{\text{dir}}\mathbf{U}^{-1}$ . Therefore,

$$\begin{aligned} \mathbf{G}_{\text{obs}} &= \mathbf{G}_{\text{dir}} + \mathbf{G}_{\text{indir}} \\ &\stackrel{(a)}{=} \mathbf{G}_{\text{dir}} + \mathbf{G}_{\text{dir}}^3 + \mathbf{G}_{\text{dir}}^5 + \dots \\ &\stackrel{(b)}{=} (\mathbf{U}\mathbf{E}_{\text{dir}}\mathbf{U}^{-1}) + (\mathbf{U}\mathbf{E}_{\text{dir}}^3\mathbf{U}^{-1}) + (\mathbf{U}\mathbf{E}_{\text{dir}}^5\mathbf{U}^{-1}) + \dots \\ &= \mathbf{U}(\mathbf{E}_{\text{dir}} + \mathbf{E}_{\text{dir}}^3 + \mathbf{E}_{\text{dir}}^5 + \dots)\mathbf{U}^{-1} \\ &= \mathbf{U} \begin{pmatrix} \sum_{i=1}^n (\lambda_{\text{dir}}^1)^{2i-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{i=1}^n (\lambda_{\text{dir}}^N)^{2i-1} \end{pmatrix} \mathbf{U}^{-1} \\ &\stackrel{(c)}{=} \mathbf{U} \begin{pmatrix} \frac{\lambda_{\text{dir}}^1}{1-(\lambda_{\text{dir}}^1)^2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\lambda_{\text{dir}}^N}{1-(\lambda_{\text{dir}}^N)^2} \end{pmatrix} \mathbf{U}^{-1} \end{aligned} \quad (7)$$

Equality (a) is inferred from the definition of diffusion model of Eq. (5); Equality (b) follows from the eigen decomposition of matrix  $\mathbf{G}_{\text{dir}}$ ; Equality (c) makes use of the character of geometric series to compute the infinite summation on the assumption of  $|\lambda_{\text{dir}}| < 1$ .

Actually, we have got the eigen decomposition of  $\mathbf{G}_{\text{obs}}$  from Eq.7. If we use  $\mathbf{E}_{\text{obs}}$  to represent the eigenvalues of  $\mathbf{G}_{\text{obs}}$ , where

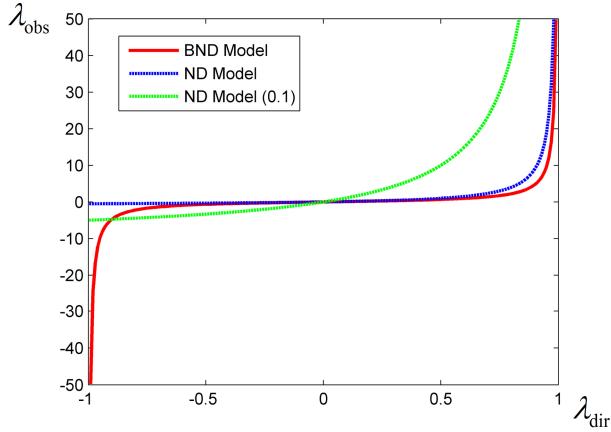
$$\mathbf{E}_{\text{obs}} = \begin{pmatrix} \lambda_{\text{obs}}^1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_{\text{obs}}^N \end{pmatrix} \quad (8)$$

We get the relationship between  $\lambda_{\text{obs}}$  and  $\lambda_{\text{dir}}$ :

$$\lambda_{\text{obs}}^i = \frac{\lambda_{\text{dir}}^i}{1 - (\lambda_{\text{dir}}^i)^2}, \quad \forall 1 \leq i \leq N \quad (9)$$

Then we solve the quadratic Eq.9 with  $\lambda_{\text{dir}}^i$  unknown and derive the nonlinear eigenvalues filter, as modeled in Eq.10, to remove the transitive noise (if  $\lambda_{\text{obs}}^i = 0$ , then  $\lambda_{\text{dir}}^i = 0$ ).

$$\lambda_{\text{dir}}^i = \frac{-1 + \sqrt{1 + 4(\lambda_{\text{obs}}^i)^2}}{2\lambda_{\text{obs}}^i} \quad (10)$$



**Fig. 1.** Plot of BND model:  $\lambda_{\text{obs}} = \frac{\lambda_{\text{dir}}}{1 - (\lambda_{\text{dir}})^2}$ , and ND models:  $\lambda_{\text{obs}} = \frac{\lambda_{\text{dir}}}{1 - \lambda_{\text{dir}}}$  and

$$\lambda_{\text{obs}} = \frac{\lambda_{\text{dir}}}{\alpha(1 - \lambda_{\text{dir}})} (\alpha = 0.1)$$

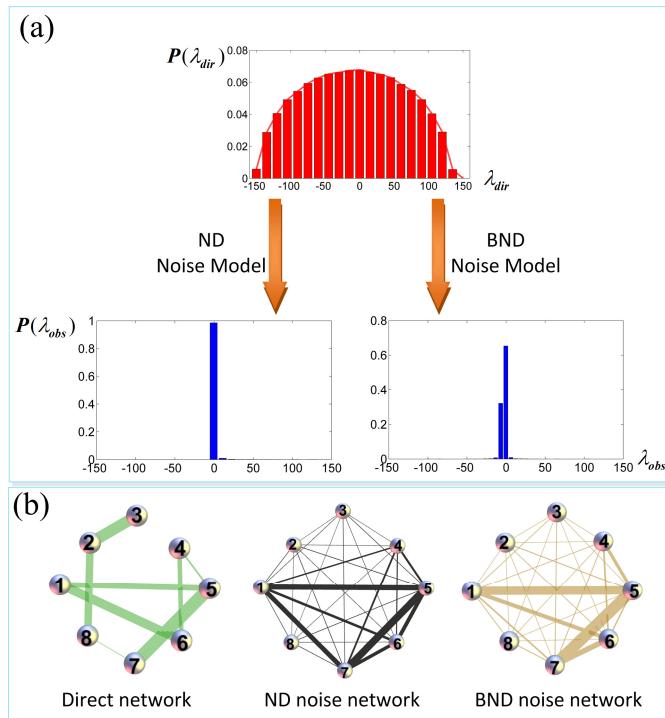
By doing this, the balanced distribution of eigenvalues can be kept, and we can infer  $\forall \lambda_{\text{obs}} \in \mathbf{R}$  from  $\lambda_{\text{dir}}$  in the range of  $|\lambda_{\text{dir}}| < 1$  (Fig.1). But ND needs a parameter  $\alpha$  to enlarge the range of  $\lambda_{\text{obs}}$ . That is why there is no further need of scaling parameters in BND.

## 4 Experimental Results

We compare the difference and similarity of BND noise model with ND's on simulated data, and test the general performance of BND on three kinds of real-world network, which are protein residue contact network, gene regulatory network and social co-authorship network.

### 4.1 Simulated Experiments

To intuitively examine both the difference and similarity of the noise models between ND and the proposed BND, we did two trial experiments on simulated data.



**Fig. 2.** (a) Eigenvalue distributions for the rebuilt noise matrices with ND and BND noise models. (b) Network topology comparison by applying ND and BND noise models on the same network.

First, we constructed a  $5000 \times 5000$  symmetric matrix containing random values drawn from the standard normal distribution. The distribution of its eigenvalues obeys Wigner's semi-circle law [17]. Then these eigenvalues are used to rebuild the noise matrix using the ND and BND noise models separately. The difference is obvious (Fig.2a): The BND noise model has maintained the balanced eigenvalue distribution while the ND noise model made almost all  $\lambda_{\text{obs}}$  positive. Similar results were observed in the real datasets.

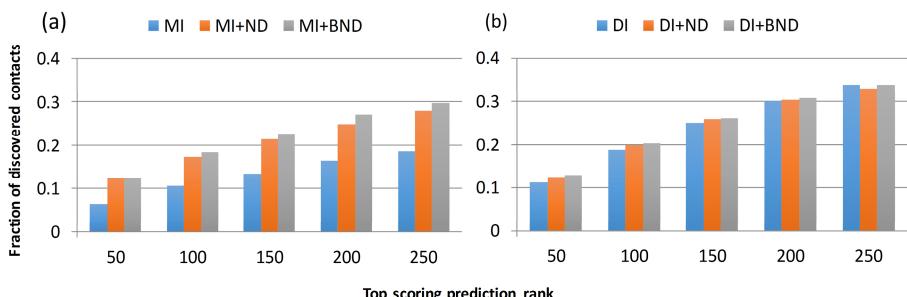
Second, we simulated an  $8 \times 8$  symmetric matrix containing pseudorandom weight values (Fig.1b). New edges were added to networks by the ND and BND noise models respectively, generating new network topologies. Two kinds of main similarity were observed: the first is that similar to ND noise model, BND's is also capable to simulate transitive noise, indicating the odd powers can cover the information in the even powers; the second is that BND noise model can also keep the strong edge weights like ND noise model, e.g. edge between nodes 5 and 7 (Fig.1c).

## 4.2 Real-Data Experiments

### 4.2.1 Protein Residue Contact Network

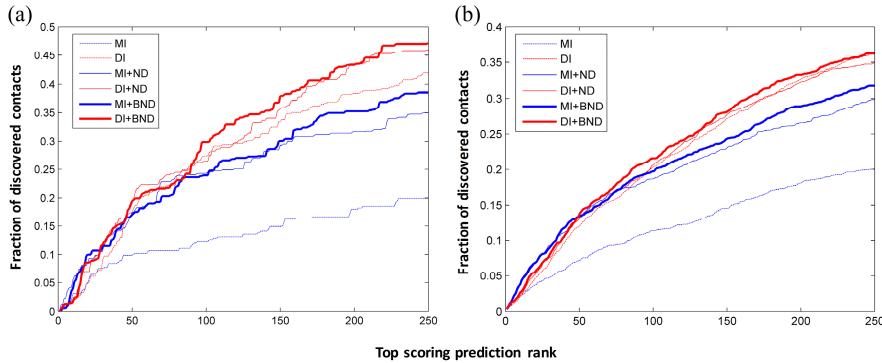
There are three steps of experiments on DI[9] dataset containing 15 proteins to test the performance of BND on residue contact prediction. First, we use MSA as input to MI [10] and DI, which are two popular methods, to predict residue-residue contact. Second, generate original contact matrix ( $\mathbf{G}_{\text{obs}}$ ) with the output from these methods. Third, BND is used to filter the transitive noise in the original contact matrix ( $\mathbf{G}_{\text{obs}}$ ).

BND improves the prediction accuracy of all the two tested methods by removing the transitive noises without the usage of an optimized tuning parameter. It is worth pointing out that the parameter of ND of this paper, which was optimized, is the same as in the original reference [4]. The average fraction of discovered residue-residue contacts is calculated in the top scoring prediction ranks from MI, DI, MI plus ND, DI plus ND, MI plus BND and DI plus BND, where the contact distance cutoff threshold for C- $\beta$  atoms is 8 Å (C- $\alpha$  in the case of Glycine). There is a consistent improvement of BND over ND in Fig.3. On average of the five ranking ranges, BND improves 6.37% on MI plus ND and 2.05% on DI plus ND. Besides, compared to MI, BND gives an improvement of 68.62%; compared to DI, BND gives an improvement of 4.11%.



**Fig. 3.** The average fraction of discovered residue-residue contacts with the threshold of contact proximity as 8 Å

In order to further test the robustness of BND, we choose two smaller thresholds of contact proximity, which are 5Å and 7Å. A smaller threshold leads to sparser contact matrices containing less false contacts. It will be more convincing if BND performs better than ND.



**Fig. 4.** The average fraction of discovered residue-residue contacts with the threshold of contact proximity as 5Å shown in part (a) and 7Å shown in part (b)

Fig.4 testifies the robustness of BND. There is consistent improvement achieved by BND compared to MI, DI and ND. Particularly, in the top 200 predictions of all contacts, BND leads to an average increase of 11.3% for all the tested proteins compared to ND plus MI. And in the top 100 predictions BND leads to an average increase of 12.9% compared to ND plus DI with the threshold of contact proximity as 5Å. Fig.4 (b) shows that BND leads to an average increase of 10.2% in the top 196 predictions and an average increase of 10.7% in the top 89 predictions with the threshold of contact proximity as 7Å.

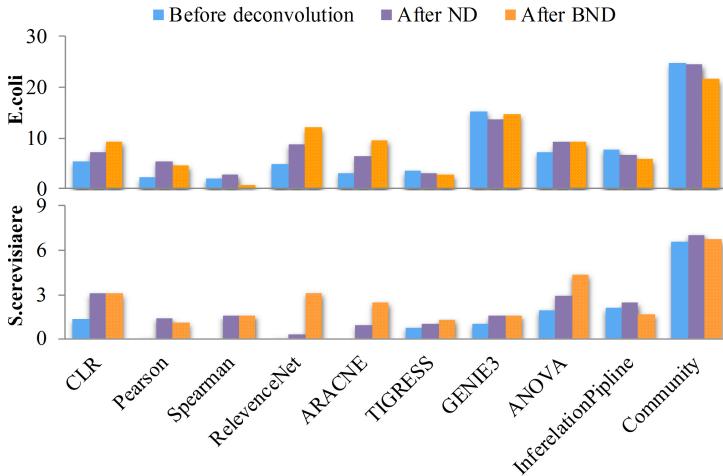
#### 4.2.2 Gene Regulatory Network Database

Apart from residue contact prediction, we test BND's ability of reconstructing gene regulatory networks, which are the same in the ND model [7]. In this experiment, the feature to predict is the gene-TF (transcription factors) and TF-TF contacts. The input matrices to ND and BND are generated from 10 popular prediction methods (Fig.5). The noise-filtered results by ND with the best performing parameter  $\beta$  equal to 0.5 and BND were compared to the experimentally verified benchmark and evaluated based on the area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) curves [14]. BND achieves 1% and 25% higher average scores of ten methods than ND on the E. coil and S. cerevisiae networks respectively.

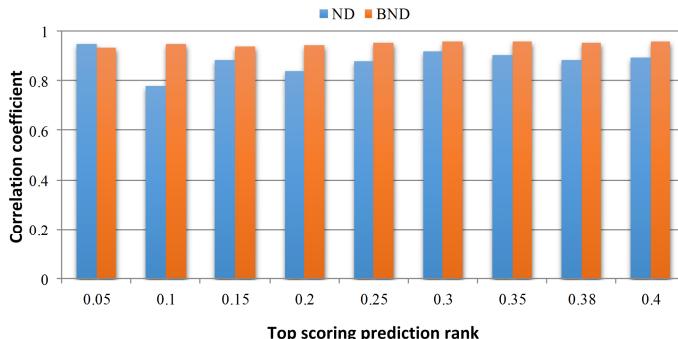
#### 4.2.3 Social Co-authorship Network

The ability of maintaining strong ties [4] in two social co-authorship networks is examined. The two networks consist of 1,589 and 16,726 nodes respectively [18], and the edges weights bigger than 0.5 are considered strong ties [4]. In the two networks, 36% and 38% of edges are strong ties. We apply BND directly and ND with the optimized  $\beta$  equal to 0.95 on these two networks. The correlation coefficients between original and re-weighted weights for the strong ties are 0.72 (BND) v.s. 0.70 (ND) on

the first network, and 0.95 (BND) v.s. 0.90 (ND) on the second network (Fig.6). These results suggest that BND performs better on maintaining original strong contacts than ND.



**Fig. 5.** The performance of BND on 10 top-scoring methods from DREAM5 [2] used by ND



**Fig. 6.** Comparation of correlation coefficients of ND and BND for maintaining strong contact weights on social co-authorship network with 16,726 nodes. The X axis represents different percentages of top ranking edges and Y axis represents the correlation coefficients between the original network and ND, BND respectively.

## 5 Conclusions

Based on the analysis of statistical behavior of eigenvalues from symmetric matrices, we propose the balanced network deconvolution (BND) method by rebuilding the noise model of network deconvolution (ND). The new noise model of transitive noise makes BND get rid of the dependence on the tuning parameter and obtain better performance.

By testing BND on three different kinds of network, we demonstrate that BND is more effective to detect amino acid contacts in protein structure network, predict TF-TF contacts in gene regulatory network and uphold strong ties in co-authorship network. Moreover, without the tuning parameter to regulate different networks used in ND, BND is a more accurate, user-friendly and general algorithm to clear the transitive correlations in observed networks.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China (No. 61222306, 91130033, 61175024), Shanghai Science and Technology Commission (No. 11JC1404800), a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (No. 201048) and Program for New Century Excellent Talents in University (NCET-11-0330).

## References

1. Newman, M.E.: The structure and function of complex networks. *SIAM Review* 45(2), 167–256 (2003)
2. Marbach, D., et al.: Wisdom of crowds for robust gene network inference. *Nature Methods* (2012)
3. Wu, S., Zhang, Y.: A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 24(7), 924–931 (2008)
4. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science* 28(6), 441–453 (2002)
5. Feizi, S., et al.: Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31(8), 726–733 (2013)
6. Jones, D.T., et al.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2), 184–190 (2012)
7. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1-2), 1–305 (2008)
8. Burger, L., van Nimwegen, E.: Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* 4, 165 (2008)
9. Weigt, M., et al.: Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* 106(1), 67–72 (2009)
10. Morcos, F., et al.: Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* 108(49), E1293–E1301 (2011)
11. Chiu, D.K., Kolodziejczak, T.: Inferring consensus structure from nucleic acid sequences. *Computer Applications in the Biosciences: CABIOS* 7(3), 347–352 (1991)
12. Faith, J.J., et al.: Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology* 5(1), e8 (2007)
13. Butte, A.J., Kohane, I.S.: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: *Pac. Symp. Biocomput.* (2000)
14. Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. *PloS One* 5(9), e12776 (2010)
15. Haury, A.-C., et al.: TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology* 6(1), 145 (2012)

16. Greenfield, A., et al.: DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS One* 5(10), e13397 (2010)
17. Küffner, R., et al.: Inferring gene regulatory networks by ANOVA. *Bioinformatics* 28(10), 1376–1382 (2012)
18. Wigner, E.P.: Random matrices in physics. *Siam Review* 9(1), 1–23 (1967)
19. Newman, M.E.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* 64(1), 016132 (2001)

# Sequence-Based Prediction of Protein-Protein Binding Residues in Alpha-Helical Membrane Proteins

Feng Xiao and Hongbin Shen\*

Institute of Image Processing & Pattern Recognition, Shanghai Jiao Tong University,  
800 Dongchuan Road, Shanghai, 200240, China  
hbshen@sjtu.edu.cn

**Abstract.** A specific number of chains form alpha-helical membrane protein complexes in order to realize the biochemical function, i.e. as gateways to decide whether specific substances can be transported across the membrane or not. However, few structures of membrane proteins have been solved. The knowledge of protein-protein binding residues can help biologists figure out how the function works and solve the 3D structures.

We present a novel, sequence-based method to predict protein-protein binding residues from primary protein sequences by machine learning classifiers. We use a support vector regression model to predict relative solvent accessibility by features based on sequences, including position specific scoring matrix, conserved score, z-coordinate prediction, second structure prediction, physical parameter and sequence length. Afterwards, combining features mentioned above with the predicted solvent accessibility, we use ensemble support vector machines to predict protein-protein binding residues. To the best of our knowledge, there is no method to predict protein-protein binding residues in alpha-helical membrane proteins. Our method outperforms MAdaBoost successfully used in predicting protein-ligand binding residues and random forest used in protein-protein binding residues from surface residues. We also assess the importance of each individual type of features. PSSM profile and conserved score are shown to be more effective to predict protein-protein binding residues in alpha-helical membrane proteins.

**Keywords:** Relative solvent accessibility, binding residues, alpha-helical membrane proteins.

## 1 Introduction

Alpha-helical transmembrane proteins (TMPs) are mostly present in the inner membranes of bacterial cells and the plasma membrane of eukaryotes. They constitute the majority of all TMPs, especially in humans. They are estimated to account for 27% of all proteins [1]. Moreover alpha-helical TMPs are often regarded as the important drug targets, i.e. G protein-coupled receptor (GPCR). Hence many efforts have been made to solve the three-dimensional structures and to understand the functions of

---

\* Corresponding author.

TMPs. However little progress has been made during past two decades, from statistical data in PDBTM database [2] by the end of 2014/05/16, there were only 2131 solved TMP structures of which 1840 are alpha-helical and 283 beta-barrel TMPs. Because of this difficulty, computational methods (template-based or ab initio) have been developed for single chain structure prediction such as Membrane-Rosetta [3] and FILM3 [4] and for several easy multi-chain complexes such as BCL::MP-Fold [5]. Accurate protein-protein binding residue prediction in membrane can help membrane complex structure prediction.

Although there are many computational methods, generally speaking, structure-based, sequenced-based methods and hybrid methods, for predicting protein-ligand binding site [6][7], only little progress have been made in protein-protein binding residue prediction in TMPs. To our knowledge the existing method proposed by Andrew J Bordner employed a Random Forest with sequence-based and structure-based features to predict the binding residues from surface residues in membrane proteins and reported the AUC of 0.75 [8]. The definition of binding residues, also using structure information, is that the surface residue has contact with another chain in the complex structure ( $< 4 \text{ \AA}$  non-H atom separation). The definition of surface residues include: (1) relative solvent accessibility surface area (RSA)  $\geq 0.2$ , (2) within the hydrophobic core of the membrane, in other words, the absolute number of the

z-coordinates predicted from the real structures are no more than  $15 \text{ \AA}$ . All the surface residues are included in the training dataset.

With the development of machine learning methods, there have been many sequence-based methods using artificial neural networks (ANNS) and support vector machines (SVMs) to predict membrane protein structure information, i.e. relative solvent accessibility (RSA) [9][10]. Although structure-based method has proven effective in protein-protein binding residues prediction, there still exists several problems needed to solve:

First, by the end of 2014/05/16, there were 101245 structures in PDB database, of which 2131 are TMPs and 1840 are alpha-helical TMPs. However the number of sequences grow rapidly contrast to the real structure considering the homology influence. So given a sequence of membrane proteins, if its real structure is not available, this structure-based method is not able to do the prediction.

Second, the existing method only predicts binding residues from surface residues in membrane proteins, in other words, before predicting protein-protein binding residues in TMPs we need to know whether the residues are surface ones or not.

In view of the above-mentioned two problems, we proposed a sequence-based protein-protein binding residue predictor for entire membrane proteins. First, we constructed a relative solvent accessibility predictor for TMP complexes with support vector regression (SVR) models. Second, protein conserved matrix (both PSSM and rate4site), predicted secondary structure matrix, predicted z-coordinate matrix, and predicted relative solvent accessibility matrix consist of the final feature set; considering the imbalance of positive (unbinding residues) and negative (binding residues) samples in our experiments, under-sampling technique was used to balance the dataset, afterwards, ensemble SVM was chosen to train the final model.

## 2 Material and Methods

### 2.1 Benchmark Dataset of Alpha-Helical Membrane Protein Complex Structures

In order to predict solvent accessibility of both single- or multiple-chain in membrane proteins, we used the same dataset as originally used in MPRAP [10]. In this dataset, the sequence identity cutoff was set to 20% and length cutoff 0.9, fragments, low-resolution structures and structures with second structure or membrane boundary problems were excluded. Thus there are 52 complexes including 80 chains in the final dataset. In order to avoid high homology in different folds, chains from the same super family were put in the same fold. The dataset was finally divided into 5 folds in advance. This dataset was also used as a benchmark dataset to predict protein-protein binding residues. It is available at [http://mprap.cbr.su.se/dataset\\_MPRAp\\_feb2010.fa](http://mprap.cbr.su.se/dataset_MPRAp_feb2010.fa). All the results showed in this paper are calculated after 5-fold cross-validation.

### 2.2 Calculation of Relative Solvent Accessibility

In this study, the RSA of each residue was calculated by Naccess 2.1.1 [11]. In our experiments we set the probe size  $1.4\text{ \AA}$  for that 2.0 did not perform well and the combination would bring error when calculating RSA intramembrane and outside separately. During the calculation for RSA of complexes, all chains in the complex were included.

### 2.3 Definition of Binding Residues

In Andrew J Bordner's work, the definition of binding residues was that (1) relative solvent accessibility surface area (SASA)  $\geq 0.2$  and the residues lied in the membrane core; (2) residues in one chain had contact with another chain in the complex structure. Contacts were defined that the atom-atom (except H-atom) distance between different residues was less than  $4\text{ \AA}$ . However Arne Elofsson's definition was that (1) relative solvent accessibility surface area (rSASA) was lower than a certain cutoff in protein complexes; (2) SASA was not lower than that certain cutoff in single-chain protein. The cutoff was set to 0.25 in Arne's work.

In our work, we define that residues in one chain contact with another chain in the complex structure are binding residues.

### 2.4 Feature Extraction

In previous work, several common features have been used successfully in the field of either solvent accessibility prediction or binding residue prediction. In this paper, we extracted 6 types of sequence-based features, including position specific scoring matrix (PSSM), conserved score by rate4site (R4S), z coordinate predictions, predicted secondary structure (SS) information, representative physical parameters (PP) and sequence length.

**Position Specific Scoring Matrix.** (PSSM) is generated by PSI-BLAST [12] to search against the UniRef90 database with 3 iterations and an E-value cutoff of 0.00001. All elements in PSSM are normalized by the following logistic function.

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

where x is the original score.

**Conserved Score** is generated by Rate4Site [13] from the multiple sequence alignment (MSA). According to Arne Elofsson et.al's work, exposed residues evolved slowly and were considered to be conserved, and buried residues evolved rapidly and were considered to be active. Thus a conclusion that the relative substitution rate was almost linearly related to the solvent accessibility in membrane protein complexes was obtained. The conserved score of each residue is normalized by subtracting the average score and dividing by the standard deviation.

**Z-coordinate Prediction** is generated by Zpred [14]. Zpred predicted the absolute z-coordinate and few predicted numbers were no more than 25 Å based on our statistics, so we normalized the predictions by dividing 25. Then the normalized numbers were added into the final feature set.

**Second Structure Prediction** is generated by PSIPRED [15]. Each residue in the sequence got the possibilities of three classes (coil (C), helix (H) and strand (E)). In our experiments, we take the three possibilities directly as the input features rather than converting to binary numbers.

**Representative Physical Parameters and Sequence Length** are residue-based features from statistical data. Representative physical parameters included a steric parameter, hydrophobicity, volume, polarity, isoelectric point, helix probability, strand probability, average accessible surface area (ASA), charge, acidity, occurrence, and average mass of twenty common amino acids. In addition to features mentioned above, sequence length was added in the feature set.

## 2.5 Using Sliding Windows to Include Neighborhood Information into Feature Set

Previous studies have indicated that the use of sliding windows can include more useful information and thus improve the prediction accuracy, i.e. second structure and relative solvent accessibility prediction [10]. In this study, we used sliding windows to cover neighborhood information in 4 types of features: (1) position specific scoring matrix (PSSM), (2) evolution rate, (3) z-coordinate prediction, (4) predicted second structure information. In our work, we found that the window size set to 9 seemed to be optimal.

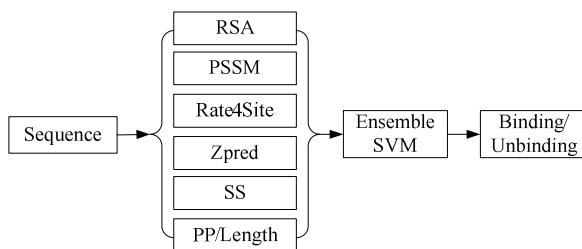
## 2.6 Prediction of Solvent Accessibility

In this section, in order to predict relative solvent accessibility a simple SVR model was used with the combination of 6 types of features. These features include: (1) PSSM; (2) Second structure prediction; (3) conserved score calculated by Rate4Site; (4) z-coordinate prediction calculated by Zpred; (5) representative physical parameters; (6) sequence length. From (1) to (4), these 4 types of features are extracted using a sliding window of length 9. (5) and (6) these two are residue-based that the sliding window is not necessary.

Afterwards, the predicted real value-RSA was also added into the binding residue-specific feature set using a sliding window, the length is set to 9.

## 2.7 Ensemble Classifier Approach to Predict Protein-Protein Binding Residues with Support Vector Machines

In order to predict binding residues, we used ensemble classifiers with support vector machines to predict membrane protein-protein binding residues from sequence information only. In our training dataset, 4629 binding residues were defined as negative samples and 16789 unbinding residues as positive samples. The ratio of positive and negative samples is about 3.6. To balance the dataset the under-sampling approach is used that positive samples with the same number of negative samples were randomly selected. Afterwards, in order to reduce the impact of under-sampling, we introduced ensemble classifiers. L different models were generated using SVMs followed by each samples, the prediction results are probabilities rather than binaries. Thus we added all the L predictions together and then divided by L.



**Fig. 1.** The flowchart of protein-protein binding residue prediction

Figure 1 illustrates the flowchart. There are 7types of features used in training dataset: (1) PSSM; (2) second structure prediction; (3) conserved score; (4) solvent accessibility prediction; (5) Z-coordinate prediction; (6) physical parameters; (7) sequence length. For a given sequence, we combined all the features together as the input to the ensemble SVM models. Finally, we got the predicted probability of binding and unbinding. In the training procedure, a certain cutoff was select to maximize the Matthews correlation coefficient. By using that certain threshold, real-valued probability was transformed to binary states (binding and unbinding).

### 3 Results and Discussions

#### 3.1 Performance of Relative Solvent Accessibility Prediction

In section 2, the predicted RSA was added to the feature set so that the performance of RSA prediction is very important. The mean absolute error (MAE) and Pearson correlation coefficient (CC) of our predictor outperforms MPRAP.

**Table 1.** Performance of different input types of features

Features	MAE	CC	MCC
MPRAP	18.202	0.583	0.470
MPRAP+SS	18.159	0.589	0.467
MPRAP+SS+length	18.136	0.593	0.472
MPRAP+SS+length+para	18.013	<b>0.599</b>	<b>0.478</b>

In Table 1, MPRAP represents three features (PSSM rate4site and Zpred) used in that method, the results (MAE: 18.202 and CC: 0.583) run in local is comparable to that reported in literature (MAE: 18.4 and CC: 0.58). Mattheus correlation coefficient (MCC) was calculated by transforming the predicted real values into binary states using a cutoff. The cutoffs were optimized to maximize MCC. We found that our predictor only added three features (SS length and parameters) improved MAE CC and MCC by 0.187 0.016 and 0.008 respectively when compared with MPRAP.

#### 3.2 Comparison with Other Methods

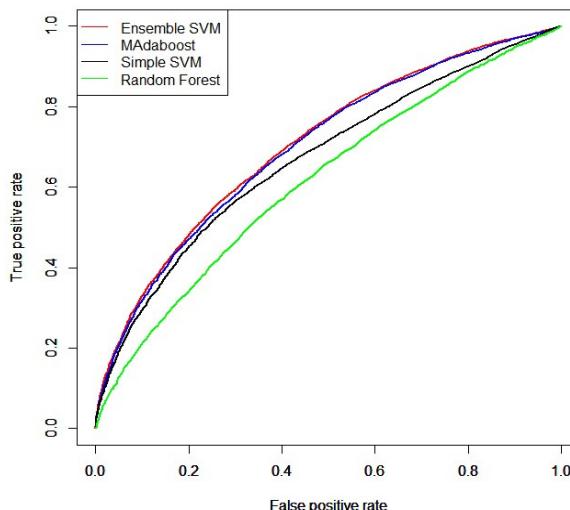
To the best of our knowledge, no work has been done to predict protein-protein binding residues in a-helical membrane proteins. Andrew J Bordner used random forests (RFs) to predict binding sites from surface residues in both a-helical and b-barrel membrane proteins. In order to do comparison, we use different methods, including (1) the MAdaBoost method used in TargetS to predict protein-ligand binding sites, (2) a simple SVM model to validate the effectiveness of the under-sampling method, (3) random forests used in Bordner's method.

In Table 2, specificity sensitivity accuracy and MCC are threshold-based, so we select the optimal threshold to maximize the MCC value. AUC is used to examine the predicted probabilities. Our method is shown to outperform other methods that it improves the AUC and MCC by 0.006 and 0.006 when comparing with the MAdaBoost method used in TargetS. Also our method achieves the best sensitivity and accuracy among all the methods. However the specificity performs not very well comparing with simple and random forests. MAdaBoost performs very in protein-ligand binding sites prediction. In our experiments, this method performs slightly worse than our method. In MAdaBoost, the base classifier is SVM. In order to evaluate the error of each base classifier, an independent dataset extracted from the training dataset is set as the evaluation dataset. So the number of samples used to train the model in MAdaBoost is less than our predictor. Considering that the number of samples is not very large, MAdaBoost is expected to achieve not very accurate result. Simple SVM and random forest achieve the high specificity and low sensitivity. These two methods

trained models in the original dataset directly. The imbalanced dataset could affect the performance of the prediction. The predictions prefer to the major class. In order to maximize the MCC value, the threshold is adjusted near the major class. This explains the high specificity and low sensitivity. Figure 2 shows the receiver operating characteristic curves of these four methods.

**Table 2.** Comparison between different methods

Methods	AUC	SPE	SEN	ACC	MCC
Ensemble SVM	0.705	0.450	0.812	0.734	0.251
MAdaBoost	0.699	0.467	0.794	0.723	0.245
Simple SVM	0.668	0.690	0.573	0.598	0.217
Random Forest	0.617	0.597	0.572	0.577	0.139

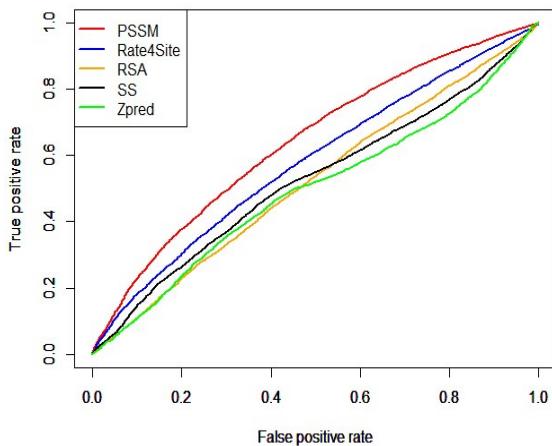


**Fig. 2.** ROC curves for different methods

### 3.3 Effectiveness of Individual Types of Input Features

In this section, we will describe the effectiveness of different types of input features in our experiments. Table 3 shows the performance of different inputs and the corresponding ROC curves are shown in Figure 3. Among the listed 5 types of features, PSSM outperforms others for that it achieves the highest AUC and MCC values, followed by Rate4Site and SS. It is expected that PSSM has proved to be the most important feature to predict protein-ligand binding residues and solvent accessibility and so on by using sequence-based methods. Rate4Site is used to calculate conserved

score and generated from the multiple sequence alignment (MSA), it performs slightly worse than PSSM and better than SS. These three features make a majority of contribution to the final prediction. RSA is obtained by our predictor directly, achieves the AUC value of 0.52 and MCC value of 0.033. By adding RSA prediction and z-coordinate prediction into feature dataset, the final results improve a little.



**Fig. 3.** ROC curves for individual types of input features

**Table 3.** Performance for individual type of features

Features	AUC	SPE	SEN	ACC	MCC
PSSM	0.641	0.478	0.719	0.667	0.174
Rate4Site	0.580	0.680	0.439	0.491	0.100
RSA	0.520	0.387	0.651	0.594	0.033
SS	0.525	0.629	0.451	0.490	0.067
Zpred	0.497	0.695	0.360	0.432	0.048

## 4 Conclusion

We developed a novel sequence-based predictor to predict protein-protein binding residues in a-helical membrane proteins. Our predictor used under-sampling methods to balance the dataset and ensemble SVMs to get the final model and outperforms other methods. We hope that our predictor would have application in guiding the experiments of solving 3-dimensional structures in membrane protein complexes.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China (No. 61222306, 91130033, 61175024), Shanghai Science and Technology Commission (No. 11JC1404800), a Foundation for the Author of National

Excellent Doctoral Dissertation of PR China (No. 201048), and Program for New Century Excellent Talents in University (NCET-11-0330).

## References

1. Almén, M.S., Nordström, K.J.V., Fredriksson, R., et al.: Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology* 7(1), 50 (2009)
2. Kozma, D., Simon, I., Tusnády, G.E.: PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Research* 41(D1), D524–D529 (2013)
3. Yarov-Yarovoy, V., Schonbrun, J., Baker, D.: Multipass membrane protein structure prediction using Rosetta. *Proteins: Structure, Function, and Bioinformatics* 62(4), 1010–1025 (2006)
4. Nugent, T., Jones, D.T.: Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis [J]. *Proceedings of the National Academy of Sciences* 109(24), E1540–E1547 (2012)
5. Weiner, B.E., Woetzel, N., Karakaş, M., et al.: BCL: MP-fold: folding membrane proteins through assembly of transmembrane helices. *Structure* 21(7), 1107–1117 (2013)
6. Chen, K., Mizianty, M.J., Kurgan, L.: ATPsite: sequence-based prediction of ATP-binding residues. *Proteome Sci.* 9(suppl. 1), S4 (2011)
7. Yu, D., Hu, J., Yang, J., et al.: Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering (2013)
8. Bordner, A.J.: Predicting protein-protein binding sites in membrane proteins. *BMC Bioinformatics* 10(1), 312 (2009)
9. Adamczak, R., Porollo, A., Meller, J.: Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins: Structure, Function, and Bioinformatics* 56(4), 753–767 (2004)
10. Illergård, K., Callegari, S., Elofsson, A.: MPRAP: An accessibility predictor for α-helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinformatics* 11(1), 333 (2010)
11. Hubbard, S.J.T.J.: NACCESS, Computer program. Department of Biochemistry and Molecular Biology 1, 1–2 (1993), <http://wolf.bi.umist.ac.uk/unix/naccess.html>
12. McGuffin, L.J., Bryson, K., Jones, D.T.: PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405 (2000)
13. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402 (1997)
14. Granseth, E., Viklund, H., Elofsson, A.: ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. *Bioinformatics* 22(14), e191–e196 (2006)
15. Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T.: Comparison of site-specific rate-inference methods: Bayesian methods are superior. *Mol. Biol. Evol.* 21, 1781–1791 (2004)

# Robust Voice Activity Detection Using the Combination of Short-Term and Long-Term Spectral Patterns<sup>\*</sup>

Yingwei Tan and Wenju Liu

National Laboratory of Pattern Recognition,  
Institute of Automation, Chinese Academy of Sciences,  
Beijing 100190, China  
[{ywtan,lwj}@nlpr.ia.ac.cn](mailto:{ywtan,lwj}@nlpr.ia.ac.cn)

**Abstract.** In this paper, we present a robust voice activity detection (VAD) algorithm using the combination of short-term and long-term spectral patterns. We analyze the benefit of short-term and long-term spectral patterns, respectively, when applied to robust VAD. Based on the analysis, we find the combination of short-term and long-term spectral patterns can be used to achieve a higher VAD accuracy than one of them only in noisy environments. We evaluate its performance under four types of noises and six types of signal-to-noise ratio (SNR) conditions. Compared with standard VAD schemes, the evaluation almost demonstrates promising results with the proposed scheme being comparable or favorable over the whole test set for various criterions of the VAD evaluation.

**Keywords:** short-term spectral patterns, long-term spectral patterns, robust voice activity detection, peak-valley difference.

## 1 Introduction

Being an important module in many applications, VAD has attracted a lot of attention in the research community over the last few decades. Many different algorithms have been proposed and the main concern of these approaches is robustness of the algorithm. The difference between most of the previous methods is the features used. Many existing VAD use features that depend on energy [1]. Some algorithms use a combination of zero-crossing rate (ZCR) and energy [2]. These algorithms do not work well in low SNRs. In [3] it is suggested that spectral peaks (SP) of audio frames are good measure for discriminating vowel sounds from other sounds including non-speech against the noisy environments even in severe noise conditions. However, for the long non-speech segment, because of the noise effect, the non-speech frames are prone to be detected as speech frames wrongly, especially when the vowel-like noise is presented. More

---

\* This research was supported in part by the China National Nature Science Foundation (No.91120303, No.61273267, No.90820011 and No.90820303).

other algorithms use more than one feature to detect speech [4, 5]. In [4], four short-term features, including spectral peaks, are used for robust VAD efficiently, but they don't always complement each other. All of the above mentioned features are typically computed from the signal along short-term analysis frames (usually 20 ms long), based on which VAD decisions are taken at each frame. In contrast to the use of the frame-level features, the use of long-term spectral divergence (LTSD) for VAD is described in [6–8]. It also shows robustness for VAD tasks. But, in the long-term analysis, the effect of adjacent frames results in wrong decision on short non-speech segments between speech segments.

In this paper, we incorporate short-term and long-term spectral patterns for robust VAD. The spectral peaks, as the short-term features, which are used for compensating for the effect of adjacent frames, namely improving the correct detection of the short non-speech segment properly, while the long-term spectral divergences, as the long-term features, which are applied for recompensing the lack of the short features, which results in the incorrect detection for the long non-speech segment. The method is compared to the most representative standards for voice activity detection such as the ITU G.729 [9] and adaptive multi-rate (AMR) [10]. The experimental results show substantial improvements.

In Section 2 the proposed robust voice activity detection algorithm are described in detail. In Section 3 we describe the speech databases, the VAD task setups, and experimental results evaluating our algorithm. Finally the conclusions are drawn.

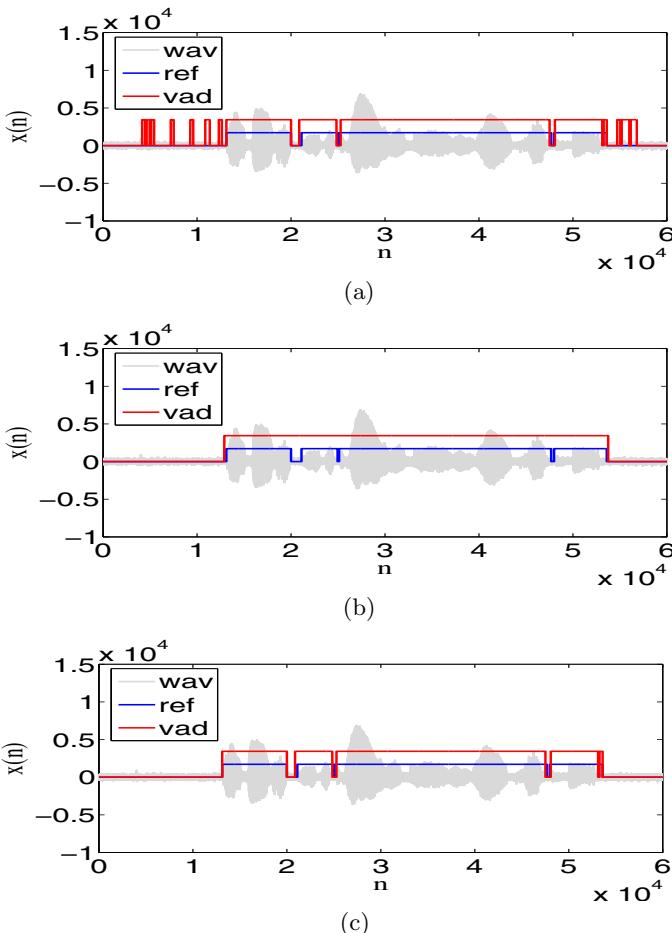
## 2 The Proposed Voice Activity Detection Algorithm

In [3], positions of spectral peaks are the most important factor in discriminating vowel sounds from others. The problem of voice activity detection is mapped to vowel sound detection. Vowel sounds have distinctive spectral peaks that are robust to various corruptions. Figure 1 (a) provides the results of an example of the operation of this VAD on an utterance of the TIMIT corpus [11].

In [7], a robust VAD algorithm has been proposed for improving speech detection robustness in noisy environments. The VAD is based on the estimation of the long-term spectral envelope and the measure of the spectral divergence between speech and noise. The decision threshold is adapted to the measured noise energy. Figure 1 (b) shows the LTSD VAD result of the same utterance.

Inspired from the observation of Figure 1 (a) and Figure 1 (b), the head or the tail of an utterance are commonly long non-speech segments. In noisy conditions, There exist some false results for the long non-speech segments through the algorithm described in [3], while the LTSD VAD algorithm proposed in [7] don't detect short pause in long speech segments, on account of the effect of adjacent frames. This paper proposes a combinational approach to improve the performance of VAD algorithm in presence of various noises and SNR conditions. Figure 1 (c) shows the results of the proposed algorithm, namely the logical conjunction of the two algorithms, for the same utterance. The results indicate that the two algorithms can complement each other. But they all prevent non-speech frames from being detected as speech frames, so this method

results in low speech hit-rate slightly. For obtaining more accurate detection, we can adjust the parameters related to the VAD threshold appropriately.



**Fig. 1.** The VAD output of an utterance from the TIMIT corpus. (a) VAD using spectral peaks, (b) VAD using LTSD, (c) The proposed VAD.

The proposed approach works with two stages. First, the training phase of the proposed approach is described with the following steps:

- i. A set of vowel segments is extracted from labeled training data.
- ii. Each segment is divided into 20 ms length frames with 10 ms overlap.
- iii. Every frame applied the fourier transform.
- iv. The average spectrum of each sound is calculated.

v. The average spectrum of vowel sound are grouped using k-means clustering algorithm.

vi. Peak detection is applied on each resulted cluster centroid.

vii. The peak signature vectors of vowel sounds, which is a binary vector containing the peak positions of a vowel sound, are extracted and stored for future reference.

In the second stage, the proposed VAD algorithm is described below:

i. Hamming windows of duration 20 ms with 10 ms between frames is applied for the noisy speech signal.

ii. Using a short-time Fourier transform, we get the short-time spectrum magnitude  $X(m, k)$ , where  $m$  and  $k$  represent the frame and frequency indices ( $k = 0, 1, \dots, NF - 1$ ).

iii. The spectrum  $X(m, k)$  is processed by means of a  $(2N + 1)$ -frame window. The  $N$ -order long-term spectral divergence between speech and noise is computed by

$$LTSD(m) = 10 \log_{10} \left( \frac{1}{NF} \sum_{k=0}^{NF-1} \frac{LTSE^2(m, k)}{N^2(k)} \right), \quad (1)$$

where the  $N$ -order long-term spectral envelope is defined as  $LTSE(m, k) = \max \{Y(m + l, k)\}_{l=-N}^{l=+N}$ , and  $N(k)$  is the noise spectrum, which is estimated using the first 10 frames and is updated by

$$N(m, k) = \begin{cases} \alpha N(m-1, k) + (1-\alpha) N_K(k) & \text{if speech} \\ N(m-1, k) & \text{otherwise} \end{cases}, \quad (2)$$

where  $N_K$  is the average spectrum over a  $K$ -frame neighbourhood

$$N_K(k) = \frac{1}{2K+1} \sum_{j=-K}^{j=-K} X(m+j, k). \quad (3)$$

Here if  $LTSD(m) - offset > \gamma_1$  and  $PVD_S <= \gamma_2$ ,  $K = 0$ , or else  $K = 3$ . The approach can eliminate interference from the adjacent frames for the accurate noise estimation. Simultaneously, the relevance of the frame to the vowel family  $V$ , namely the peak-valley difference (PVD), is calculated with the following measure

$$PVD_S = \max_{S \in V}(PVD(X, S)), \quad (4)$$

$$PVD(X, S) = \frac{\sum_{k=0}^{NF-1} (X(k) \times S(k))}{\sum_{k=0}^{NF-1} S(k)} - \frac{\sum_{k=0}^{NF-1} (X(k) \times (1 - S(k)))}{\sum_{k=0}^{NF-1} (1 - S(k))}, \quad (5)$$

where  $S$  is the spectral peaks signature of the current frame. Frequency indices are ignored for simplicity here.

iv. If  $LTSD(m) - offset > \gamma_1$  and  $PVD_S > \gamma_2$ , we mark the current frame as speech, otherwise, mark it as silence.

There are two decision threshold in the proposed approach. They are  $\gamma_1$  and  $\gamma_2$ , respectively. The decision threshold  $\gamma_1$  is adapted to the measured noise energy  $E$  by

$$\gamma_1 = \begin{cases} \gamma_1^0 & \text{if } E \leq E_0 \\ \frac{\gamma_1^0 - \gamma_1^1}{E_0 - E_1} E + \gamma_1^0 - \frac{\gamma_1^0 - \gamma_1^1}{1 - \frac{E_1}{E_0}} & \text{if } E_0 < E < E_1, \\ \gamma_1^1 & \text{if } E \geq E_1 \end{cases} \quad (6)$$

where  $E_0$  and  $E_1$  are the energies of the background noise for the cleanest and noisiest conditions can be determined examining the speech database being used, and Optimal parameters  $\gamma_1^0$  and  $\gamma_1^1$  for clean and high noise conditions, respectively , are defined. As for  $\gamma_2$ , using the first 10 frames, we obtain

$$\gamma_2 = \max_{t \in 1 \dots 10} \left( \max_{S \in V} (PVD(X_t, S)) \right) + \theta, \quad (7)$$

where the optimal value of  $\theta$  can be found on a set of evaluation speech data so that the total performance of the algorithm on the evaluation data maximizes.

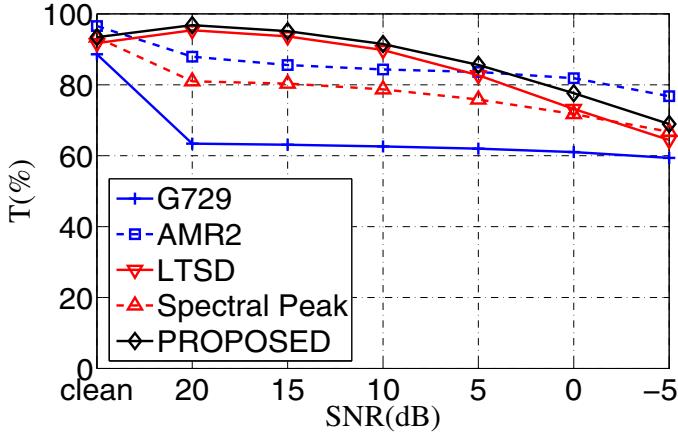
### 3 Experiments and Results

Performances of the proposed VAD in noisy environments including the white, babble, factory, and street noises at different SNRs are investigated for comparison.

According to [12], the choice of test data is important in VAD evaluation, since it is simple to optimize a particular scheme for a small test set. Here, we use the entire TIMIT test corpus [11] consisting of 168 individual speakers of eight different dialects, each speaking ten phonetically balanced sentences. Then noise of each category is added at six different SNR levels (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB) to all 1680 sentences. During obtaining the reference labels or performing the VAD, sentences are concatenated for each speaker, silence is inserted between sentences, the total amount of silence in the sentence set of each speaker is constrained to be 60% of samples, and there is a result for the concatenated sentences of each speaker.

The proposed VAD is evaluated in terms of the ability to discriminate between speech and pause periods at different SNR levels. There are two common metrics for VAD performance evaluation, which are speech pause hit-rate (HR0) and speech hit-rate (HR1). In order to obtain a better metric for comparing two different VAD algorithms, the mean of HR0 and HR1 is applied as the final evaluation metric ( $T$ ).  $T$  is averaged for the entire set of noises, for clean conditions and SNR levels ranging from 20 to -5 dB with a 5 db step. The parameters used for the proposed VAD are:  $N = 6$ ,  $\gamma_1^0 = 15$ ,  $\gamma_1^1 = -2.4$ ,  $offset = 5$ ,  $E_0 = 26.87$ ,

$E_1 = 101.30$ ,  $\alpha = 0.95$ ,  $\theta = 2$ . Figure 2 compares the VADs in terms of the average  $T$ . The proposed approach obtains the best behavior with a 87.02% T average value of all conditions, while G.729, AMR2, the spectral peak (SP) based VAD, and LTSD yield 65.73%, 85.25%, 78.25%, and 84.41%, respectively.



**Fig. 2.** The mean of non-speech hit-rate (HR0) and speech hit-rate (HR1)

Then, we follow the testing strategy proposed by [13]. This evaluation is performed through five different parameters reflecting the VAD performance.

- 1) CORRECT: Correct decisions made by the VAD.
- 2) FEC (front end clipping): Clipping due to speech misclassified as noise in passing from noise to speech activity.
- 3) MSC (mid speech clipping): Clipping due to speech misclassified as noise during a speech region.
- 4) OVER (over hang): Noise interpreted as speech due to the VAD flag remaining active in passing from speech activity to noise.
- 5) NDS (noise detected as speech): Noise interpreted as speech within a silence period.

FEC and MSC are indicators of true rejection, while NDS and OVER are indicators of false acceptance. CORRECT parameter indicates the amount of correct decisions made. Thus all four parameters FEC, MSC, NDS, OVER should be minimized and the CORRECT parameter should be maximized to obtain the best overall system performance.

Table 1 shows that on the average, the proposed scheme is better than LTSD in terms of CORRECT score for white (2.68%), babble (6.87%), factory (6.72%), and street (2.60%). Compared to the AMR2, the proposed VAD is better in terms of CORRECT score for babble (10.38%) and factory (14.28%), and worse for white (5.02%) and street (1.77%).

**Table 1.** CORRECT, FEC, MSC, OVER, and NDS averaged over clean conditions and all SNR levels for four noises as obtained by six VAD schemes - G.729, AMR2, SP, LTSD, Proposed

Method	White noise					Babble noise				
	CORRECT	FEC	MSC	OVER	NDS	CORRECT	FEC	MSC	OVER	NDS
G.729	55.73	<b>0.08</b>	<b>0.2</b>	20.61	23.38	50.39	<b>0.01</b>	<b>0.03</b>	27.31	22.27
AMR2	<b>92.99</b>	0.68	3.37	2.13	<b>0.84</b>	76.15	0.15	0.54	9.05	14.10
SP	75.56	0.36	4.03	5.02	15.02	76.14	0.59	4.98	4.62	13.67
LTSD	85.29	2.75	8.29	2.41	1.26	79.66	0.09	0.50	14.79	4.95
Proposed	87.97	2.27	6.08	<b>1.77</b>	1.91	<b>86.53</b>	0.75	6.08	<b>2.19</b>	<b>4.45</b>
Factory noise						Street noise				
Method	CORRECT	FEC	MSC	OVER	NDS	CORRECT	FEC	MSC	OVER	NDS
G.729	67.91	<b>0.16</b>	<b>0.73</b>	5.04	26.17	64.63	<b>0.01</b>	<b>0.04</b>	8.06	27.26
AMR2	70.78	0.23	1.79	3.82	23.37	<b>93.74</b>	0.21	0.63	3.32	<b>2.10</b>
SP	80.84	0.77	7.99	1.30	9.10	74.09	0.19	2.26	4.29	19.16
LTSD	78.34	0.26	1.57	6.15	13.67	89.37	0.12	0.62	7.75	2.15
Proposed	<b>85.06</b>	1.29	9.32	<b>1.01</b>	<b>3.31</b>	91.97	0.29	2.98	<b>1.79</b>	2.96

## 4 Conclusions

This paper has shown the proposed algorithm for the overall VAD system performance, particularly, in noisy conditions. The effectiveness of the proposed algorithm are discussed in detail. The presented VAD is based on the combination of short-term and long-term spectral patterns. Its advantages lie in making full use of the characteristics of two state-of-the-art algorithms and making them work in harmony. The main flaw of this approach is that it focus on preventing non-speech frames from being detected as speech frames and may result in low speech hit-rate. However we can adjust the parameters related to the threshold and reduce the unfavourable factor. As a result, we obtain more robust VAD in noisy environments.

## References

1. Evangelopoulos, G., Maragos, P.: Speech event detection using multiband modulation energy. In: Proceedings of INTERSPEECH, pp. 685–688 (2005)
2. Kotnik, B., Kacic, Z., Horvat, B.: A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. In: Proceedings of INTERSPEECH, pp. 197–200 (2001)
3. Yoo, I.-C., Yook, D.: Robust voice activity detection using the spectral peaks of vowel sounds. ETRI Journal 31(4) (2009)
4. Moattar, M.H., Homayounpour, M.M., Kalantari, N.K.: A new approach for robust realtime voice activity detection using spectral pattern. In: Proceedings of ICASSP, pp. 4478–4481 (2010)
5. Soleimani, S.A., Ahadi, S.M.: Voice activity detection based on combination of multiple features using linear/kernel discriminant analyses. In: Proceedings of ICTTA, pp. 1–5 (2008)

6. Ramirez, J., Segura, J.C., Benitez, M., de la Torre, A., Rubio, A.: A new adaptive long-term spectral estimation voice activity detector. In: Proceedings of EUROSPEECH, pp. 3041–3044 (2003)
7. Ramirez, J., Segura, J.C., Benitez, C., De La Torre, A., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. *Speech Communication* 42(3), 271–287 (2004)
8. Ramirez, J., Segura, J.C., Benitez, C., de La Torre, A., Rubio, A.: Voice activity detection with noise reduction and long-term spectral divergence estimation. In: Proceedings of ICASSP (2004)
9. Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., Petit, J.-P.: ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications. *IEEE Communications Magazine* 35(9), 64–73 (1997)
10. ETSI, Voice activity detector(VAD) for Adaptive MultiRate(AMR) speech traffic channels, ETSI EN 301 708 Recommendation (1999)
11. Garofolo, J.S., et al.: Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. In: National Institute of Standards and Technology (NIST), Gaithersburgh, MD, vol. 107 (1988)
12. Sarikaya, R., Sarikaya, R., Hansen, J.H.L., Hansen, J.H.L.: Robust speech activity detection in the presence of noise. In: Proceedings of ICSLP (1998)
13. Beritelli, F., Casale, S., Cavallaero, A.: A robust voice activity detector for wireless communications using soft computing. *IEEE Journal on Selected Areas in Communications* 16(9), 1818–1829 (1998)

# Speech Emotion Recognition Based on Coiflet Wavelet Packet Cepstral Coefficients

Yongming Huang<sup>1,2</sup>, Ao Wu<sup>1,2</sup>, Guobao Zhang<sup>1,2</sup>, and Yue Li<sup>1,2</sup>

<sup>1</sup> School of Automation, Southeast University, Nanjing 210096, China

<sup>2</sup> Key Laboratory of Measurement and Control of Complex Systems of Engineering,  
Ministry of Education

**Abstract.** A wavelet packet based adaptive filter-bank construction method is proposed for speech signal processing in this paper. On this basis, a set of acoustic features are proposed for speech emotion recognition, namely Coiflet Wavelet Packet Cepstral Coefficients (CWPCC). CWPCC extends the conventional Mel-Frequency Cepstral Coefficients (MFCC) by adapting the filter-bank structure according to the decision task; Speech emotion recognition system is constructed with the proposed feature set and Gaussian mixture model as classifier. Experimental results on Berlin emotional speech database show that the Coiflet Wavelet Packet is more suitable in speech emotion recognition than other Wavelet Packets and proposed features improve emotion recognition performance over the conventional features.

**Keywords:** Speech emotion recognition, Coiflet Wavelet packets Cepstral Coefficients (CWPCC), Acoustic features.

## 1 Instruction

Speech emotion recognition shows broad application prospects in various fields. For example, it can be employed in automatic telephone systems to help process phone calls according to perceived urgency [1], in healthcare service to help diagnose depression and suicide risk [2], in the design of interactive movies and online games where natural responds are required from users [3], and in intelligent automobile systems to monitor drivers mental state and ensure safety [4].

In recent years, wavelet packet (WP) is efficient in providing flexible and adaptive frequency band division methods [5], and is a prominent technique for quasi-periodic and non-stationary signal processing, such as speech processing. In this paper, the problem of constructing proper tree-structured WP basis that adapt the frequency partition solution to the emotion classification task is explored. On this basis, novel Coiflet WP-based acoustic features are proposed for speech emotion classification.

## 2 Wavelet Packet Transform

In this section, a brief introduction of WP is provided. For excellent expositions, readers are referred to [5, 6].

For a discrete input signal  $x(n)$  as explained in [5], to compute WP coefficients, we should first associate to  $\bar{x}(t)$ , approximated at resolution  $d_0^0 = a_0$  with decomposition coefficients  $a_0(n)$  that satisfy

$$x(n) = f_s^{1/2} a_0(n) \approx \bar{x}(n \cdot f_s^{-1}) \quad (1)$$

where  $f_s$  denotes the sampling rate. Then a pair of conjugate mirror filters (CMF)  $h(n)$  and  $g(n)$  is applied on  $a_0(n)$  which decomposes the signal into approximation and detail components. The recursive decomposition can be represented by an admissible binary tree structure, where each node has either zero or two children.

Let  $\mathcal{T} = \{(0,0), (1,0), (1,1), \dots, (J,0), \dots, (J, 2^J - 1)\}$  denote a WP admissible binary tree with depth  $J$ .  $(j, p) \in \mathcal{T}$  is labeled by its depth ( $j$ ) in the tree and the number of nodes ( $p$ ) on its left at depth  $j$ . By applying the CMF on WP coefficients  $d_j^p$ , node  $(j, p)$  is split into two child nodes  $(j+1, 2p)$  and  $(j+1, 2p+1)$ , with corresponding WP coefficients  $d_{j+1}^{2p}$  and  $d_{j+1}^{2p+1}$ , respectively.

$$d_{j+1}^{2p}(n) = \sum_{r=-\infty}^{+\infty} h(r-2n) d_j^p(r) \quad (2)$$

$$d_{j+1}^{2p+1}(n) = \sum_{r=-\infty}^{+\infty} g(r-2n) d_j^p(r) \quad (3)$$

For the root node  $(0, 0)$  the associated WP coefficients are  $d_0^0 = a_0$ .

By iterating the splitting of WP tree nodes in a particular way, an admissible binary tree structure is obtained. From a filter-bank point of view, the tree structure represents a frequency band partition method. Various WP filterbank structures can be achieved according to different signal analysis purposes. For the speech emotion recognition task, our goal is to generate a frequency band division method according to the distribution of emotion information along frequency axis, and acquire emotion-related acoustic features by multi-channel filtering with the WP filter-bank. The issue of determining an optimal frequency band division method is interpreted as the WP tree pruning problem and will be investigated in the following section.

### 3 WP Tree Pruning Problem

The set of leaf nodes of  $\mathcal{T}$  is denoted as  $\mathcal{L}(\mathcal{T})$  and the set of internal nodes is denoted as  $\mathcal{I}(\mathcal{T})$ . We denote a full rooted binary tree with root  $v_{\text{root}} = (0, 0)$  as  $\mathcal{T}_{\text{full}}$ . Among the sub-trees rooted at  $v$  with  $k$  leaf nodes, the one maximizes the adopted criterion is denoted by  $\mathcal{T}_v^k$ . We denote the left and right children of node  $v \in \mathcal{I}(\mathcal{T})$  as  $l(v)$  and  $r(v)$  respectively. The size of  $\mathcal{T}$  is defined as the number of terminal nodes, and we denote it by  $|\mathcal{T}|$ . If  $\mathcal{T}_{l(v)}^*$  and  $\mathcal{T}_{r(v)}^*$  are pruned sub-trees rooted at  $l(v)$  and  $r(v)$  respectively, let  $\llbracket v, \mathcal{T}_{l(v)}^*, \mathcal{T}_{r(v)}^* \rrbracket$  denote the pruned sub-tree rooted at  $v$  with  $\mathcal{T}_{l(v)}^*$  and  $\mathcal{T}_{r(v)}^*$  as its left and right sub-trees, respectively.

Let  $x \in V_{J_0}$  be the signal observed, and  $W_j^p x$  be the component of  $x$  in subspace  $W_j^p$ . We use  $M(j, p; x)$  as a measurement for  $W_j^p x$ , with which the discrimination power of node  $(j, p) \in \mathcal{T}$  can be calculated. In this work, the measurement is specified as signal energy, a widely adopted measurement as in [7, 8, 9].

The energy-based measurement is defined as

$$M(j, p; x) = \|W_j^p x\|^2 / \|x\|^2 \quad (4)$$

In the rest of this paper, energy-based measurement is used as the quantity supplied for discriminate power calculation at each node of the WP tree.

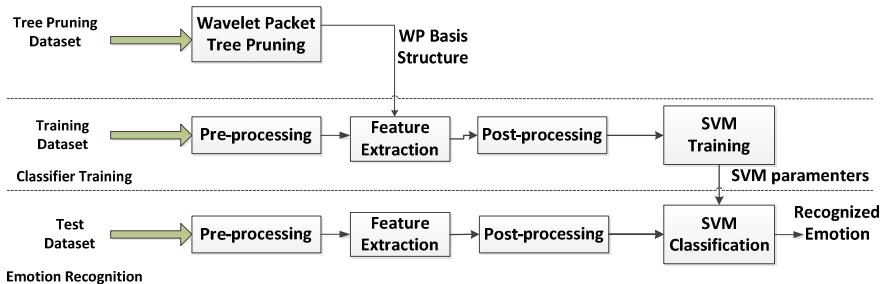
Given an additive discrimination measure  $D(\mathcal{T})$ , a simple addition is operated instead of computing the functional on the union of the nodes, therefore, a fast algorithm can be applied for the tree pruning problem.

### 4 Proposed System

In this section, we describe details of the proposed speech emotion recognition system with emphasis on the WP filter-bank based acoustic feature extraction. The following subsections give detailed description of each part of the proposed system. Framework of the proposed speech emotion recognition system is shown in Fig.1.

#### 4.1 Dataset Division

The whole emotional speech database is divided into three parts. The tree-pruning dataset, training dataset and test dataset are used for WP tree pruning, classifier training and emotion recognition respectively, and are not overlapped with each other.



**Fig. 1.** Block diagram of the proposed system

#### 4.2 Wavelet Packet Filter-Bank Construction

An optimal filter-bank structure is obtained using the tree-pruning algorithm. The optimal WP filter-bank structure is then applied on the training and test samples to calculate WP-based acoustic features.

With the fast tree-pruning algorithm, a sequence of WP admissible trees with different number of leaf nodes is obtained, and correspondingly the set of filter-bank structures with different number of sub-bands. The obtained WP filter-bank structures are then used to calculate emotion-discriminative acoustic features from original speech signal.

#### 4.3 Pre-processing

Before feature extraction, conventional speech signal processing operations including pre-emphasis, frame blocking and windowing are performed first. The speech signal is first pre-emphasized by a high-pass FIR filter  $1 - 0.9375z^{-1}$  to spectrally flatten the signal and make it less susceptible to finite precision effects later in the signal processing [10]. The pre-emphasized speech signal is then blocked into frames of  $K$  samples with an overlap of  $K'$  between adjacent frames. Here we use  $K=256$  and  $K'=K/2$ . And each individual frame is multiplied by a Hamming window to reduce ripples in the spectrum.

#### 4.4 Feature Extraction

A type of wavelet packet features is proposed for the speech emotion recognition task, namely the Coiflet Wavelet Packet Cepstral Coefficients (CWPCC).

The first feature set CWPCC is implemented in a similar way to the conventional MFCC features. First, the windowed speech frame is passed through a filter-bank obtained from WP tree pruning step, and then the sub-band log-energy coefficients are calculated and frequency ordered, followed by the Discrete Cosine Transform (DCT) to de-correlate the log filter-bank energy. Finally, the first DCT coefficients together with the log-energy of the frame are adopted, with the first and second order derivatives appended to the feature vector.

## 4.5 Classifier

Support vector machine (SVM) is adopted for speech emotion classification in this paper. For each emotion class a SVM is trained with the training dataset. The implementation of the SVM classifier is provided by a publicly available Matlab toolbox named LIBSVM Matlab Toolbox [11].

# 5 Experiments

## 5.1 Emotional Speech Database and Experimental Setup

The proposed speech emotion recognition system is evaluated on the Berlin emotional speech database [12], which contains 7 simulated emotions (anger, boredom, disgust, fear, joy, neutral and sadness). In this paper, six emotions (no disgust) with a sum of 489 utterances are used for the classification task. The disgust emotion is discarded because the number of disgust utterances is fairly limited. 20% of the database is randomly selected to form the tree-pruning dataset, and we apply 5-fold cross validation on the remaining 80% utterances to assess the classification performance.

## 5.2 Experimental Results

To evaluate performance of the proposed feature, a set of experiments are conducted and the results are presented below. The maximum level of WP decomposition is set to  $J = 5$ , and the dimensionality of WPPC feature is 39, which is the same with conventional MFCC feature. As mentioned above, the de-correlation ability of WP filter-banks depends on the type and order of the CMF. In this work, the wavelet packet filters are used as CMFs, which are orthogonal filters with compact support [6]. Different wavelet packets are used to generate WP filter-bank structures and compared. Daubechies wavelets of orders 32 and 40 (denoted as db32, db40), coiflet wavelets of orders 1, 2, 3, 4 and 5 (denoted as coif 1, coif 2, coif 3, coif 4, coif5), symlet wavelets of orders 8 and 45 (denoted as sym8 and sym45).

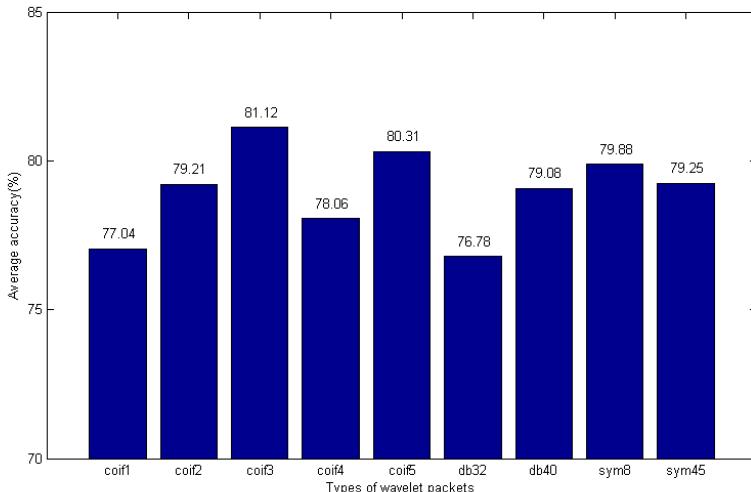
WP filter-bank structures generated by coif 3 achieved the highest accuracy rate in speech emotion recognition compared with other wavelet packets.

The coiflets wavelet has some advantages, such as orthogonal, compact support, near symmetric and so on. Wavelet coifN is more symmetric than dbN. The bracing length of coifN is same as db3N and its vanishing moment numbers is same as db2N. So coiflets wavelet achieved the best performance.

Each row in the confusion matrix represents a percentage of the speech emotion sample to be identified as different types of emotions. The diagonal elements are the emotion recognition accuracy of various types of emotions. From Table 1, we can see CWPCC has strong ability to identify anger, boredom, fear and sadness. However, joy is easily confused with anger and fear.

We also do the experiments to compare the speech emotion recognition performance between CWPCC and conventional features. Prosody features include pitch frequency and logarithmic frame energy while the acoustic features include the

first, the second and the third formant. From Table 2, we can see the CWPCC improved the emotion recognition accuracy of six emotions.



**Fig. 2.** Emotion recognition rates adopted different wavelet packets

**Table 1.** Confusion Matrix with Coif3

Emotion	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	<b>89.22%</b>	0.00%	2.94%	7.84%	0.00%	0.00%
Boredom	0.00%	<b>84.62%</b>	1.54%	0.00%	6.15%	7.69%
Fear	1.82%	1.82%	<b>76.36%</b>	7.27%	5.45%	7.27%
Joy	28.07%	0.00%	14.04%	<b>57.89%</b>	0.00%	0.00%
Neutral	0.00%	4.76%	3.17%	0.00%	<b>88.89%</b>	3.17%
Sadness	0.00%	16.00%	0.00%	0.00%	2.00%	<b>82.00%</b>
<b>Average recognition rate:</b> 81.12%						

**Table 2.** The comparision between different features

Features	Recognition Rate (%)						
	Anger	Boredom	Fear	Joy	Neutral	Sadness	Average
CWPCC	<b>91.18</b>	<b>87.69</b>	<b>74.55</b>	<b>64.91</b>	<b>90.48</b>	<b>86.00</b>	<b>83.67</b>
Prosody Features	71.65	70.37	52.17	39.44	45.57	66.13	59.10
Acoustic Features	81.10	59.26	43.48	46.48	54.43	70.97	61.55

### 5.3 Experiments with the CWPCC Feature Set

Coiflet filter of order 3 was used to generate wavelet packets. Experiments were conducted with the number of Cepstral coefficients fixed to 12 (i.e. feature dimension is 39) and the number of sub-bands varied from 15 to 25.

It is also worth noting that the WP filter-bank structures with sub-band number 17 and 21 generally perform better among different feature configurations. This result is consistent for different tree pruning criteria and filter orders. The corresponding frequency partition solution gives us an insight into the emotion-related information distribution along frequency axis and provides us with better knowledge of humans emotion perception mechanism through acoustic signal.

A highest classification accuracy of 81.12% is achieved for the proposed CWPCC feature set, with Coiflet filter of order 3.

For the WPCC feature with the best classification performance, the corresponding pruned wavelet packet tree structure and frequency partition solution of [0,8kHz] is illustrated in Fig.2.

## 6 Conclusion and Future Work

In this paper we explored the wavelet packet based acoustic feature extraction approach for speech emotion recognition.

A type of short-time acoustic features was proposed based on the obtained coiflet WP filter-bank structure, namely Coiflet Wavelet Packet Cepstral Coefficients (CWPCC). The CWPCC feature is calculated following the conventional Mel-frequency Cepstral analysis paradigm. Finally, a speech emotion recognition system was built and experiments were carried out on the Berlin emotional speech database to evaluate the proposed feature sets. Experimental results demonstrate the superiority of the proposed feature set over conventional feature.

Future work also includes investigating more effective WP tree pruning scheme which, for example, not only takes the discriminating ability as tree pruning criterion, but also has the size of WP tree controlled in a reasonable range. Apart from this, seeking for robust feature representation is also considered as part of the ongoing research, as well as efficient classification techniques for automatic speech emotion recognition.

**Acknowledgements.** This work was supported by open Fund of the Key Laboratory of Measurement and Control of CSE (School of Automation, Southeast University), Ministry of Education (no. MCCSE2013B03), Jiangsu Province Natural Science Foundation (no.BK20140649).

## References

1. Morrison, D., Wang, R.L., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. *Speech Comm.* 49(2), 98–112 (2007)
2. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M.: Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering* 47(7), 829–837 (2000)

3. Caponetti, L., Buscicchio, C.A., Castellano, G.: Biologically inspired emo-tion recognition from speech. *Eurasip Journal on Advances in Signal Processing*
4. Malta, L., Miyajima, C., Kitaoka, N., Takeda, K.: Multimodal estimationof a driver's spontaneous irritation. In: *Intelligent Vehicles Symposium*, pp. 573–577. IEEE (2009)
5. Stephane, M.: *A Wavelet Tour of Signal Processing*, 3rd edn. Academic Press, Burlington (2009)
6. Daubechies, I.: *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia (1992)
7. Pavez, E., Silva, J.F.: Analysis and design of wavelet-packet cepstral co-efficients for automaticspeech recognition. *Speech Comm.* 54(6), 814–835 (2012)
8. Saito, N., Coifman, R.R.: Local discriminant bases. In: *SPIE 2303, Mathematical Imaging:Wavelet Applications in Signal and Image Processing*, pp. 2–14 (1994)
9. Silva, J., Narayanan, S.S.: Discriminative wavelet packet filter bank selection for pattern recognition. *IEEE Trans. Signal Process.* 57(5), 1796–1810 (2009)
10. Rabiner, L., Juang, B.-H.: *Fundamentals of Speech Recognition*. Prentice-Hall, New Jersey (1993)
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3), 1–27 (2011)
12. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: *Proceeding INTERSPEECH 2005, ISCA*, pp. 1517–1520 (2005)

# Text Detection in Natural Scene Images Leveraging Context Information

Runmin Wang, Nong Sang\*, Changxin Gao, Xiaoqin Kuang, and Jun Xiang

School of Automation, Huazhong University of Science and Technology,  
Wuhan, China, 430074  
{runminwang, nsang, cgao, kxqkuang}@hust.edu.cn,  
xiangjun.0511@163.com

**Abstract.** In this paper, we propose a method leveraging context information for text detection in natural scene images. Most of the existing methods just utilize the hand-engineered features to describe the text area, but we focus on building a confidence map model by integrating the candidate appearance and the relationships with its adjacent candidates. Three layers of filtering strategy is designed to judge the category of the text candidates, which can remove abundant non-text regions. In order to retrieve the missing text regions, a context fusion step is performed. Finally, the remaining connected components (CCs) are grouped into text lines and are further verified, and then the text lines are broken into separate words. Experimental results on two benchmark datasets, i.e., ICDAR 2005, ICDAR 2013, demonstrate that the proposed approach has achieved the competitive performances with the state-of-the-art algorithms.

**Keywords:** Text detection, context information, confidence map, natural scene image.

## 1 Introduction

Text information plays a significant role in many applications for providing a lot of descriptive and abstract information. Many demands in real life application, e.g. object recognition, assistive navigation, scene understanding, image-based search, etc., make text detection to be a crucial task in content-based image analysis techniques. Text detection is a challenging problem for texts in natural scenes having variations of font size, colors and alignment orientation. Moreover, it is often affected by complex background, illumination changes, image distortion and degradation. Due to text detection in natural scenes plays a significant role in many applications, numerous detection methods have been proposed in recent years, and these methods are roughly classified into two categories: CC-based methods and texture-based methods.

CC-based methods base on the facts that the characters in text regions exhibit certain properties, e.g. approximately constant color, proximate pixel value,

---

\* Corresponding author.

similar stroke width, etc.. Various approaches are used to get the CCs, such as Maximally Stable Extremal Regions(MSER) [14] [21], Stroke Width Transform(SWT) [1], K-means clustering [15], etc. CC-based methods segment an image into a set of CCs, and the CCs are classified as text or background by analyzing their geometrical characteristics. CC-based methods are relatively fast, however, these methods can not segment text components accurately without prior knowledge of texts, and it is difficult to design a reliable CC analyzer for existing many non-text components. Meanwhile, this kind of approaches encounter difficulties when the text is noisy, multi-colored and textured.

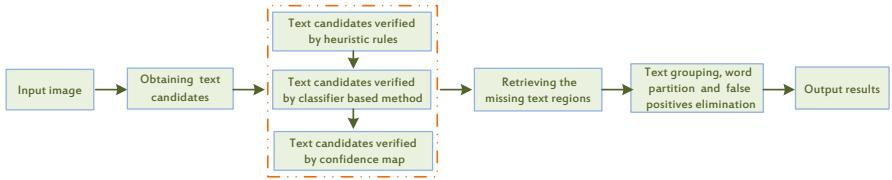
To overcome the problem of complex background, the texture-based approaches consider text as a special texture region. In the texture-based methods, various kinds features, e.g., gradient edge features [16], T-HOG [13], etc., extracted from each candidate region are fed into trained classifiers, and the text likelihood of the candidate region is estimated by Support Vector Machine(SVM) [19], Neural Networks [7] and AdaBoost [4], etc. Texture-based approaches are efficient in dealing with complex background problem. However, these methods always use trained classifier scanning the entire image with a sliding window, and many thousands of predictions are demanded. In addition, a large number of training samples are needed to train the classifier, which are more time-consuming.

Different with document images, in which text characters are normalized into simple background and proper resolutions, texts in natural scenes always confront many unfavorable situations, e.g., the variations of text font, complex background, illumination changes, image distortion and degradation, which make it challenging to recognize the texts in natural scenes. As a result, if we adopt the traditional method to deal with the scene texts, the poor recognition performance will be received. To solve this problem, different from most of the existing methods, which recognize the candidates by only using the hand-engineered features, we focus on establishing a confidence map model by integrating the candidate performance and the relationships with its adjacent candidates to improve the classification performance. In this work, the candidate appearance is evaluated by using a trained classifier based on the state-of-the-art feature descriptor, and some context relationships are integrated to boost the robustness of the classification results. Following this way, the texts can be highlighted from the background.

The main contributions in this paper mainly include: (1) A scheme has been designed by reasonably using the texture-based method and CC-based method to robustly detect texts in natural scenes. (2) A confidence map model is proposed to highlight texts from the background by integrating the candidate appearance and the context relationships with its adjacent candidates.

## 2 System Overview

The framework of our system is shown in Fig.1. Firstly, the text candidates are produced via the image binarization processing by using the graph cut inference, and then the candidates are verified by three layers of filtering strategy.

**Fig. 1.** Flowchart of the proposed algorithm

Secondly, a context fusion step is performed to retrieve the missing text regions. Finally, the text lines are further verified and broken into separate words.

### 3 Our Methodology

#### 3.1 Finding Text Candidates

The characters in text regions always exhibit certain properties, such as approximately constant color, proximate pixel value, similar stroke width, etc. In our work, we adopt the method proposed in [12] based on the fact that this method has been verified particularly suitable for scenes text binarization. This method embeds local binarization into a global optimization framework, and the global optimization problem is solved by using the graph cut inference. An important advantage of this method is not need any information about the position and size of the text in an image.

In [12], the Niblack binarization method has been used in binarization processing. Different with the foreground pixels and the background pixels, some ambiguity pixels are tagged 0.5 in the binarization image. In order to clarify the binarization results, we extend the binarization method to improve the results in our work, and a judgment strategy is designed as the last step added to their pipelines. For each pixel tagged 0.5, the average value with  $3 \times 3$  neighborhood will be calculated. If the average value is bigger than a certain threshold, the pixel will be tagged 1, otherwise it will be tagged 0 (the binarization result is shown in Fig.3(b)). In order to describe the algorithm conveniently, we illustrate this question with some formulas as follows:

$$\hat{G}(i, j) = \frac{1}{9} \times \sum_{m,n=-1}^1 G(i+m, j+n) \quad i = 2, \dots, M-1, j = 2, \dots, N-1 \quad (1)$$

$$\begin{cases} G(i, j) = 0, & \text{if } \hat{G}(i, j) \leq 0.78 \\ G(i, j) = 1, & \text{if } \hat{G}(i, j) > 0.78 \end{cases} \quad (2)$$

Where the  $G(i, j)$  is pixel value, and the  $\hat{G}(i, j)$  is the average value of the pixel with  $3 \times 3$  neighborhood at the coordinate position  $(i, j)$ . The  $M, N$  represent the number of row and column of the image respectively. Since the texts in natural scenes can be either darker than background or lighter than background, we perform the aforementioned processing twice.

### 3.2 Text Candidates Verification

The text candidates can be obtained through the aforementioned image binarization processing, however, abundant non-text regions will also be produced. In order to remove these false positives, three layers filtering strategy is designed at this stage. The heuristic rules remove the candidates most unlikely the texts firstly, and then taking into account the time consumption of integrating the relationships of the adjacent candidates in confidence map, the classifier based method is performed to remove the false positives, which contributes to reducing the time consumption of the confidence map processing.

**False Positives Elimination by Heuristic Rules.** In order to distinguish the text candidates from the background, the geometric property of single component is utilized. Considering excessive parameters may reduce the robustness of the algorithm, only three heuristic rules are adopted, i.e., width, height and area of the CCs. In our work, the three parameter thresholds were learned on the ICDAR 2005 annotated training set<sup>1</sup>. For the text, the range of variation of these parameters as follows:

$$\begin{aligned} \text{CC\_height} &\in [9, 0.70 \times \text{img\_H}], \quad \text{CC\_width} \in [5, 0.72 \times \text{img\_W}] \\ \text{CC\_area} &\in [40, 0.15 \times \text{img\_H} \times \text{img\_W}] \end{aligned} \quad (3)$$

Where  $\text{img\_H}$ ,  $\text{img\_W}$  are the height and the width of the input image,  $\text{CC\_height}$ ,  $\text{CC\_width}$  and  $\text{CC\_area}$  are the height, width and area of the CC.

**False Positives Elimination by Classifier Based Method.** As aforementioned analysis, it is difficult to design a reliable CC analyzer to deal with the complex background, some false positives can not be removed in the previous processing. Since there may be existing adhesion between adjacent characters after the binarization processing, two kinds of classifiers, i.e., two-class text line recognizer and multi-class character recognizer, are designed to handle the adhesion and the non-adhesion phenomenon respectively. Note that, the existence of adhesion between the adjacent characters is judged by the aspect ratio of the connected component in our work. An important advantage of this hierarchical identification strategy adopted in our work is to reduce the texture distortion caused by the image normalization processing.

In our work, we judge the candidates without adhesion by a multi-class character recognizer, which has been trained by 13791 training samples from the ICDAR 2005 training dataset<sup>2</sup>, and the training samples are normalized to size  $48 \times 48$ . It is worth noting that we just need to distinguish the text character from the non-text interferences, so we have merged some similar character categories, such as "P, p", "S, s", "B, 8", etc. As a result, there are 47 kinds of training samples (uppercase, lowercase letters, digits and non-text) in our training dataset.

---

<sup>1</sup> Available at: <http://graphics.cs.msu.ru/en/research/projects/msr/text>

<sup>2</sup> Available at

<http://algoval.essex.ac.uk/icdar/Datasets.html#RobustReading.html>

In order to deal with the case of candidates with adhesion, we train the classifier based on two kinds of training samples (text line and non-text line), in which the training samples are normalized to size  $144 \times 48$ . Due to the proposed algorithm is evaluated based on the two benchmark datasets, i.e. ICDAR 2005 dataset, ICDAR 2013 dataset, we build the training sets to train the classifiers respectively for fairness. There are 2338 positive samples and 2709 negative samples are collected from the ICDAR 2005 training datasets, 3633 positive samples and 3768 negative samples are collected from the ICDAR 2013 training datasets<sup>3</sup>.

In our work, the features used for describing the texture characteristic of the candidates are the Histogram of Oriented Gradients (HOG) in [3], which has been widely accepted as one of the best features in object detection and recognition field. In our work, we train the aforementioned classifiers based on SVM, and each candidate region is divided into cells of  $8 \times 8$  pixels and each group of  $2 \times 2$  cells is integrated into a block. Meanwhile, each cell consists of 9 orientation bins. After the aforementioned processing, the candidates unlikely texts will be removed, (the result is shown in Fig.3(c).

Different with document images, in which text characters are normalized into simple background and proper resolutions, recognizing the text in natural scenes is a challenging topic. Many negative factors, e.g., the variations of text font, complex background, illumination changes, image distortion and degradation, will affect the recognition performance. Although some texts may be removed with low text likelihood, it's highly unlikely that the texts in the same text line are removed simultaneously. Accordingly, we can retrieve these missing texts by using the method introduced in the Sec.3.3.

**False Positives Elimination by Confidence Maps.** So far, most of the existing methods just use some hand-engineered features, e.g. gradient edge features [16], T-HOG [13], etc., to describe the text area. As aforementioned introduction, some negative factors affect the recognition performance, and some background outliers may be recognized as text with high text likelihood. In order to remove the false positives in further, we focus on establishing a confidence map model by integrating the candidate performance and the relationships with its adjacent candidates. The candidates with low confidence value below a certain threshold will be removed.

In our experiments, we assume that the text lines in natural scenes usually are horizontal or slightly tilted. We respectively regard every remaining connected component as seed candidate firstly, and then several similarity, e.g. height similarity, color similarity, stroke width similarity, etc., between the seed candidate and its adjacent candidates in horizontal direction will be calculated. As shown in Fig.2(b), the text candidate  $R$  highlighted with red background is seed candidate, and the other candidates in horizontal direction, i.e.  $C, A, P, A, R, K$ , are the adjacent candidates of the seed candidate  $R$ . The calculation process of the confidence map is shown in Fig.2(c), and the confidence map result of the seed

---

<sup>3</sup> Available at

<http://dag.cvc.uab.es/icdar2013competition/?ch=2&com=downloads>

candidate  $R$  is shown in Fig.2(d). We describe the confidence map model from the view of mathematics as follows:



**Fig. 2.** Processing of confidence map

$$\begin{aligned} Char\_tc(i) = & \alpha \times Char\_hogc(i) + \beta \times Char\_hc(i, j) \\ & + \gamma \times Char\_swc(i, j) + \psi \times Char\_rgbc(i, j) \end{aligned} \quad (4)$$

$$Char\_hc(i, j) = \left( \frac{1}{N} \right) \sum_{j=1}^N \left( 1 - \frac{|Char\_h(i) - Char\_h(j)|}{\max(Char\_h(i), Char\_h(j))} \right) \quad (5)$$

$$Char\_swc(i, j) = \left( \frac{1}{N} \right) \sum_{j=1}^N \left( 1 - \frac{|Char\_sw(i) - Char\_sw(j)|}{\max(Char\_sw(i), Char\_sw(j))} \right) \quad (6)$$

$$Char\_rgbc(i, j) = \left( \frac{1}{N} \right) \sum_{j=1}^N \left( 1 - \sum_{R, G, B} \sum_{k=1}^b \left( \frac{|h(i, k) - h(j, k)|}{\max(h(i, k), h(j, k))} \right) \right) \quad (7)$$

For the  $i$ -th candidate, the  $Char\_tc(i)$  presents the total confidence value.  $Char\_hogc(i)$  present the text likelihood judged by the aforementioned trained classifiers.  $Char\_hc(i, j)$ ,  $Char\_swc(i, j)$  and  $Char\_rgbc(i, j)$  present the height similarity, stroke width similarity and color similarity of the  $i$ -th candidate and its  $j$ -th adjacent candidate in horizontal direction respectively.  $h(i, k), h(j, k)$  present the color histogram,  $k$  is the number of quantitative in color histogram, and the  $k = 256$  in our work. The  $i \in (1, M), j \in (1, N)$ , the  $M$  represents the total number of the remaining CCs, and  $N$  represents the total number of the adjacent candidates of the  $i$ -th candidate in horizontal direction. The  $\alpha, \beta, \gamma, \psi$  represent weights of the aforementioned four values, and in our work,  $\alpha = 1, \beta = 2, \gamma = 1, \psi = 1$ . As shown in Fig.3(d), the color of the candidates corresponding to the right colorbar represent their texts likelihood, and most of the non-texts with low confidence value below a certain threshold are removed.

To further verify the effectiveness of the confidence maps, we adopt the aforementioned ICDAR2005 annotation dataset as the baseline. The detection results are compared with the baseline by pixel level, and they are quantitatively measured by the *Recall* and the *Precision*. Note that, at this section, the two parameters are defined as follows:

$$Recall = \frac{\sum (img\_dec \cap img\_gt)}{\sum img\_gt} \quad (8)$$

**Fig. 3.** Results of text candidates verification**Table 1.** Comparative results by using confidence map

Method	Precise	Recall
before using confidence map	0.44	0.69
after using confidence map	0.60	0.66

$$\text{Precision} = \frac{\sum (img\_dec \cap img\_gt)}{\sum img\_dec} \quad (9)$$

Where, the  $img\_dec$  is the detection result, and the  $img\_gt$  is the corresponding ground truth. As shown in Table 1, we can greatly improve the *Precision* with a little costs of the *Recall* reduction.

### 3.3 Retrieving Missing Text Candidates

In order to retrieve the missing text regions, a context fusion step is performed based on the fact that the texts in natural scenes always appear in groups and text lines. The remaining candidates after the aforementioned processing are analyzed, and the adjacent candidates with similar height in the horizontal direction are grouped to form the key regions. In order to further illustrate this question, we denote the missing text candidates as  $Mr = \{mr_1, \dots, mr_K\}$ ,  $K$  is the missing candidate total number. The  $mr_i$  on behalf of the  $i$ -th missing text candidate, and the state of  $mr_i$  is defined as  $\{cs_i, cc_i, cw_i, ch_i\}$ , where the  $cs_i$  is stroke width,  $cc_i$  is the common part with corresponding key region,  $cw_i$  and  $ch_i$  is the whole width and the whole height respectively. Meanwhile, we denote the current key region as  $Kr = \{kr_1, \dots, kr_N\}$ ,  $N$  is the candidate number in the key region.  $ks_{ave}$  is the average stroke width of the candidates in this key region,  $kw_{ave}$  and  $kh_{ave}$  are the average width and the average height respectively. The missing text candidates will be retrieved in the search area, if they satisfy:

$$\begin{cases} \min(cc_i, kw_{ave}) / \max(cc_i, kw_{ave}) < T_1, & \text{when in vertical search area} \\ \min(cc_i, kh_{ave}) / \max(cc_i, kh_{ave}) < T_1, & \text{when in horizontal search area} \end{cases} \quad (10)$$

$$\min(cs_i, ks_{ave})/\max(cs_i, ks_{ave}) < T_2 \quad (11)$$

$$\min(cw_i, kw_{ave})/\max(cw_i, kw_{ave}) < T_3 \quad (12)$$

$$\min(ch_i, kh_{ave})/\max(ch_i, kh_{ave}) < T_4 \quad (13)$$

Note that, the first row of *Formula 10* is one of the requirements to find the missing text regions in vertical search area, and the second row of *Formula 10* is corresponding to the horizontal search area. We set  $T_1 = 0.5$ ,  $T_2 = T_3 = T_4 = 0.75$  in this work.

### 3.4 Text Line Formed and Segmented into Word

The adjacent candidates with similar height in horizontal direction form the text lines, and then the text lines are further verified by the off-the-shelf CNN classifier proposed in [17]. Although word segmentation is not the main issue of the scene text detection problem, we need to break the text line into separate words according to the evaluation strategy in ICDAR robust reading competition. Given a text line, we calculate the distances between adjacent CCs in the text line and obtain the average distance. The minimum word spacing distance  $T$  is estimated by *Eq 14*, and separation between them is occurred, if the distance between adjacent CCs overs  $T$ .

$$T = \alpha \times D_{ave} + \beta \quad (14)$$

Where  $D_{ave}$  is the average distance value. In our work, we empirically set  $\alpha = 1.75$  and  $\beta = 3$ .



**Fig. 4.** Examples of detection results of the proposed method

## 4 Experiment Results

In order to verify the effectiveness of the text detection scheme proposed in this paper, we carry out experiments on the two public datasets, i.e., ICDAR 2005, ICDAR 2013. Some examples of detection results of the proposed method are

shown in Fig.4, these results indicate that our system is robust against large variations in text font, color, size, and geometric distortion. However, as shown in Fig.4., our system maybe miss a few text regions with poor resolution, wrongly separate a few intra-word letters from the inter-word letters, and left over a small amount of false positives, etc., these problems will be further researched in our future work. The performance of our method is further quantitatively measured by *Precision* (P), *Recall* (R) and *f – Measure* (F) introduced in [10] [6], and the performances of our method and other state-of-the-art algorithms on the ICDAR 2005 and ICDAR 2013 databases are shown in Table 2 and Table 3. In order to distinguish the results of other methods, our results are represented with the bold font.

**Table 2.** Experimental results on the ICDAR 2005 dataset

Method	Year	P	R	F
<b>Our method</b>	—	<b>0.75</b>	<b>0.61</b>	<b>0.67</b>
Epshtain [1]	2010	0.73	0.60	0.66
Li [8]	2013	0.62	0.65	0.63
Yi [20]	2013	0.71	0.62	0.63
Hinnerk Becker (1st ICDAR 2005) [9]	2005	0.62	0.67	0.62
Quan Meng [11]	2012	0.66	0.57	0.61
Yao [18]	2007	0.64	0.60	0.61
Zhang [5]	2010	0.67	0.46	-
Ashida (1st ICDAR 2003) [10]	2003	0.55	0.46	0.50

**Table 3.** Experimental results on the ICDAR 2013 dataset [6]

Method	Year	P	R	F
USTB TexStar (1st ICDAR 2013)	2013	0.89	0.67	0.76
<b>Our method</b>	—	<b>0.75</b>	<b>0.58</b>	<b>0.66</b>
Text Detection [2]	2013	0.74	0.53	0.62
Baseline	2013	0.61	0.35	0.44
Inkam	2013	0.31	0.35	0.33

As shown in Table 2 and Table 3, our method achieved the *Precision* 0.75, *Recall* 0.61 and *f – Measure* 0.67 for the ICDAR 2005 dataset, and we can obtain the *Precision* 0.75, *Recall* 0.58 and *f – Measure* 0.66 for the ICDAR 2013 dataset respectively. Comparing our approach with the other state-of-the-art algorithms listed in the Table 2 and Table 3, our approach has achieved the competitive performances with the most state-of-the-art algorithms on ICDAR 2005 and ICDAR 2013 datasets. Specifically, it need to point out that our results are superior to the method proposed in [8], in which the context information has also been adopted. In addition, we can get the better results than the method [1], which has been certified to be one of the most efficient method to detect the scene texts.

## 5 Conclusion

In this work, a scheme is designed by reasonably using the texture-based method and CC-based method to robustly detect texts in natural scenes. Different with the most of the existing methods only by using the hand-engineered features to describe the text candidates, a confidence map model is built to strengthen the recognition performance by integrating the candidate appearance and the relationships with its adjacent candidates. In order to verify the effectiveness of the proposed text detection method, we carry out experiments on two baseline datasets, and the experimental results demonstrate that the proposed method is effective in unconstrained scene text detection.

**Acknowledgment.** The authors would like to thank the National Natural Science Foundation of China (No.61105014, No.61302137, No.61401170), and the Natural Science Foundation of Hubei Province (2013CFB403) for supporting the research.

## References

1. Epshtain, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2963–2970 (2010)
2. Fabrizio, J., Marcotegui, B., Cord, M.: Text detection in street level images. *Pattern Analysis and Applications* 16(4), 519–533 (2013)
3. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1627–1645 (2009)
4. Hanif, S., Prevost, L.: Text detection and localization in complex scene images using constrained adaboost algorithm. In: Proceeding of International Conference on Document Analysis and Recognition, pp. 1–5 (2009)
5. Zhang, J., Kasturi, R.: Text detection using edge gradient and graph spectrum. In: Proceedings of the International Conference on Pattern Recognition, pp. 3979–3982 (2010)
6. Karatzas, D., Shafait, F., Uchida, S., et al.: Icdar 2013 robust reading competition. In: Proceedings of the IEEE Conference on Document Analysis and Recognition, pp. 1484–1493 (2013)
7. Li, H., Doermann, D., Kia, O.: Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing* 9(1), 147–156 (2000)
8. Li, Y., Shen, C., Jia, W., van den Hengel, A.: Leveraging surrounding context for scene text detection. In: Proceedings of the IEEE International Conference on Image Processing, pp. 2264–2268 (2013)
9. Lucas, S.M.: Icdar 2005 text locating competition results. In: Proceeding of International Conference on Document Analysis and Recognition, pp. 80–84 (2005)
10. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: Icdar 2003 robust reading competitions. In: Proceeding of International Conference on Document Analysis and Recognition, pp. 682–687 (2003)
11. Meng, Q., Song, Y.: Text detection in natural scenes with salient region. In: Proceeding of the IAPR International Workshop on Document Analysis Systems, pp. 384–388 (2012)

12. Milyaev, S., Barinova, O., Novikova, T., Kohli, P., Lempitsky, V.S.: Image binarization for end-to-end text understanding in natural images. In: Proceeding of International Conference on Document Analysis and Recognition, pp. 128–132 (2013)
13. Minetto, R., Thome, N., Cord, M., Leite, N.J., Stolfi, J.: T-hog: An effective gradient-based descriptor for single line text regions. *Pattern Recognition* 46(3), 1078–1090 (2013)
14. Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S.: Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recognition Letters* 34(2), 107–116 (2013)
15. Shivakumara, P., Phan, T.Q., Tan, C.: A laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2), 412–419 (2011)
16. Shivakumara, P., Huang, W., Phan, T.Q., Tan, C.L.: Accurate video text detection through classification of low and high contrast images. *Pattern Recognition* 43(6), 2165–2185 (2010)
17. Wang, T., Wu, D., Coates, A., Ng, A.: End-to-end text recognition with convolutional neural networks. In: Proceeding of International Conference on Pattern Recognition, pp. 3304–3308 (2012)
18. Yao, J., Wang, Y., Weng, L., Yang, Y.: Locating text based on connected component and svm. In: Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, pp. 1418–1423 (2007)
19. Ye, Q., Huang, Q., Gao, W., Zhao, D.: Fast and robust text detection in images and video frames. *Image and Vision Computing* 23(6), 565–576 (2005)
20. Yi, C., Tian, Y.: Text extraction from scene images by character appearance and structure modeling. *Computer Vision and Image Understanding* 117(2), 182–194 (2013)
21. Yin, X.C., Yin, X., Huang, K., Hao, H.W.: Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(5), 970–983 (2014)

# Adaptive Local Receptive Field Convolutional Neural Networks for Handwritten Chinese Character Recognition

Li Chen, Chunpeng Wu, Wei Fan, Jun Sun, and Naoi Satoshi

Fujitsu Research & Development Center Co. Ltd., Beijing, 100025, China  
{chenli,wuchunpeng,fanwei,sunjun,naoi}@cn.fujitsu.com

**Abstract.** The success of convolutional neural networks (CNNs) in the field of image recognition suggests that local connectivity is one of the key issues to exploit the prior information of structured data. But the problem of selecting optimal local receptive field still remains. We argue that the best way to select optimal local receptive field is to let CNNs learn how to choose it. To this end, we first use different sizes of local receptive fields to produce several sets of feature maps, then an element-wise max pooling layer is introduced to select the optimal neurons from these sets of feature maps. A novel training process ensures that each neuron of the model has the opportunity to be fully trained. The results of the experiments on handwritten Chinese character recognition show that the proposed method significantly improves the performance of traditional CNNs.

**Keywords:** Convolutional Neural Networks (CNNs), Local Receptive Field, Handwritten Chinese Character Recognition.

## 1 Introduction

In the last several years, CNNs have achieved the state of the art in a variety of classification tasks, such as digit recognition [1], Chinese character recognition [2], human face recognition [3], human pose estimation [4], ImageNet recognition competition [5] and Speech Recognition [6]. Compared to other types of feed forward neural networks (such as Deep Believe Network or Auto-Encoder) where neurons of two adjacent layers are fully connected, CNNs enforce a local connectivity pattern that each neuron only connects with a small local subset of the neurons (called local receptive field) of the previous layer. In addition, neurons in each local receptive field across a feature map share the same set of parameters. These two properties make CNNs extremely suitable for structured data with local smoothness, such as image and audio. However, it also yields a question: how can we choose an optimal size of the local receptive field?

Intuitively, CNNs with large local receptive fields have the capability to learn complex features which may represent the raw data well, but have more parameters. Experimental evidence shows that large neural networks are fundamentally hard to learn [7]. What's more, CNNs with large local receptive fields take less advantages of

prior information. On the other side, Small local receptive fields make CNNs more compact thus easy to learn, but have less ability to represent the raw data. An ideal CNN should adaptively determine the size of local receptive field with large size on smooth regions and small size on regions possessing abundant details.

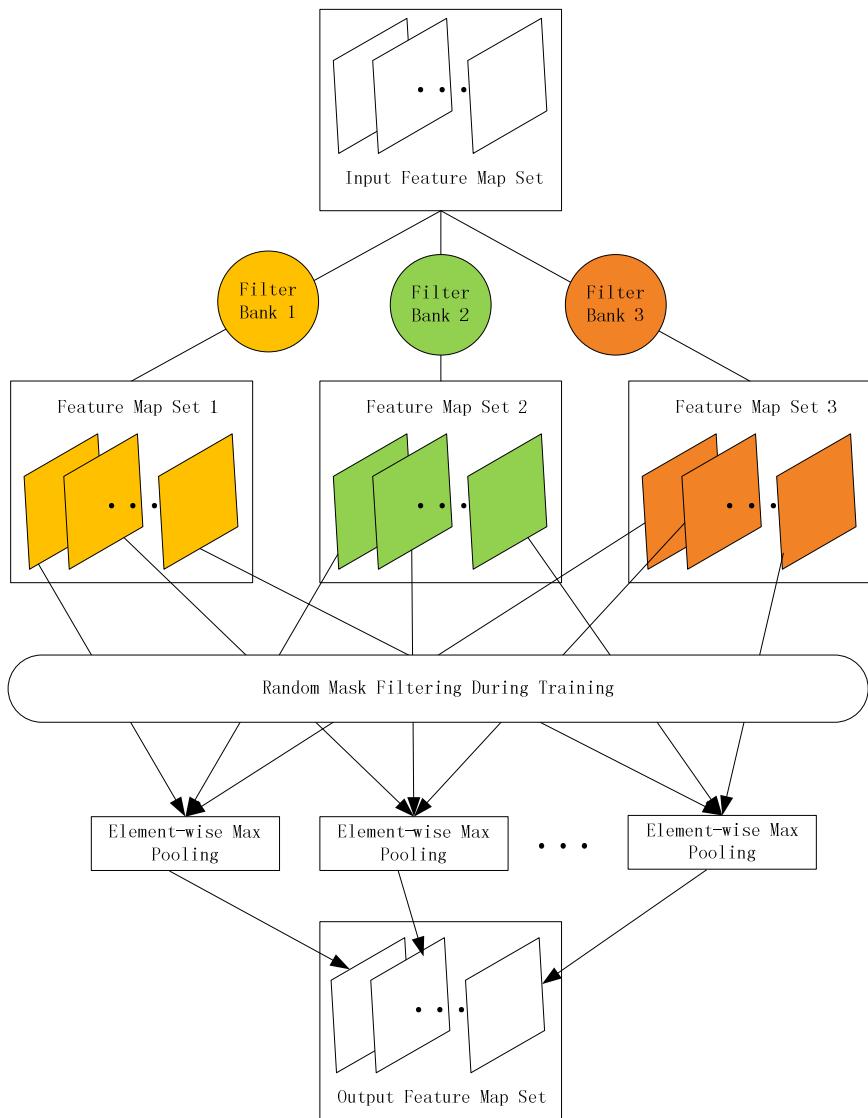
A common way to select the size of local receptive field is to enumerate all possible sizes and choose the best one. But it is labor-intensive and time-consuming. Some researchers simply choose a pretty good local receptive field size according to their experience which is typically not the best one.

Some literatures concerning this question were published in the last few years. Adam Coates et al [8] investigated the effect of receptive field size in a single-layer network using unsupervised feature learning and drew a conclusion that small receptive fields can work well and should be preferred. They also developed an algorithm to choose the local receptive field by measuring the similarity of low-level features [9]. In [10], the authors developed a method to learn the receptive fields of pooling layers. And in [11], an algorithm called collaborative receptive field learning is proposed to extract specific receptive fields from multiple images. While the above methods tackle the receptive field selection problem in different perspectives, few literatures focus on the problem of local receptive field selection in CNNs.

On the other hand, some important tricks were developed very recently which significantly improve the performance of deep neural networks (DNNs). The most promising one is dropout network developed by Hinton [12]. A dropout network is a kind of network where some neurons are randomly dropped out at each iteration of the training procedure and the activation values of corresponding neurons are scaled in the testing step. Dropout opens a door of improving the generalization ability of DNNs by injecting some kind of random noise to the model during training, and has several extensions, including dropconnect [13], Stochastic Pooling [14], maxout [15], and ALT-CNN[16]. Maxout network extends the idea of dropout by introducing the competition strategy where neurons in the same layer are grouped and only the neuron with the largest activation can pass to the next layer in each group.

Inspired by dropout and maxout strategies, we propose a local receptive field selection algorithm for CNNs. In the proposed method, kernels with different sizes are used to produce several sets of feature maps. An element-wise max pooling layer is introduced to select neurons from these sets of feature maps. To make the network fully trained, a random mask is generated for each input feature map set of the element-wise max pooling layer during training procedure, which ensures the network to be trained effectively and efficiently. We apply our method to handwritten Chinese character recognition and get a significant improvement of the performance compared to the conventional CNNs.

The rest of this paper is organized as follows. In Section 2, we describe the structure and training process of our adaptive local receptive field CNNs in detail. Section 3 gives the experimental results on handwritten Chinese character recognition. The conclusion and the future work are given in Section 4.



**Fig. 1.** The structure of modified convolutional layer

## 2 Proposed Method

Traditional CNNs typically consist of several convolutional layers, several pooling layers and one or more fully connected layer. We modify the structure of the convolutional layer by introducing different sizes of local receptive fields to produce different sets of feature maps and an element-wise max pooling layer to choose the neurons from these sets of feature maps. We also modify the training process by

adding random disturbance to the modified convolutional layer to make sure that all neurons have the opportunity to be fully trained.

## 2.1 Structure of the Modified Convolutional Layer

In traditional CNNs, a convolutional layer has many filters with the same size and produces a single set of feature maps. In the modified convolutional layer, on the contrary, a convolutional layer uses several sets of filters with different sizes to produce several sets of feature maps. The number of feature maps in each set is the same, and all feature maps have the same size. Neurons at the same position from different sets represent features produced by different local receptive fields at that position. An element-wise max pooling layer is used to select neurons at each position from all feature map sets. It compares the values of the elements in the same position across all sets and chooses the maximum. The result is then sent to the next layer.

Fig.1 shows the structure of the proposed modified convolutional layer. The input feature map set is filtered by three filter banks. Filters from different filter banks have different sizes of receptive fields. Each filter bank produces a feature map set. An element-wise max pooling layer then selects the maximum values at each location from three feature map sets.

## 2.2 Training Process

Randomly omitting some neurons in the training process is proved to be an effective way to increase the generalization ability of DNNs. We adopt this strategy by introducing a random mask filtering operation before element-wise max-pooling layer (see Fig.1). In each iteration during training procedure, a random mask is generated for every input feature map of the element-wise max-pooling layer. We generate the random mask according to the following three principles: a) at least one neuron can pass the mask at each location, ensuring that in each place there is a neuron getting trained at each iteration; b) some location should allow more than one neuron to pass the mask, ensuring that the competition mechanism works in the training process; c) each neuron has equal possibility to pass the mask. To this end, we first use a random matrix to assign each place a neuron, then, for each feature map set we generate a random matrix to allow some additional neurons to pass the mask:

$$M_k^{(1)}(i, x, y) = \begin{cases} 1, & (k-1)/N < R^{(1)}(i, x, y) < k/N \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$M_k^{(2)}(i, x, y) = \begin{cases} 1, & R_k^{(2)}(i, x, y) < t \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $M_k^{(1)}$ ,  $M_k^{(2)}$ ,  $R^{(1)}$  and  $R_k^{(2)}$  have the same shape.  $M_k^{(1)}(i, x, y)$  indicates the mask value in the position  $(x, y)$  of the  $i$ th feature map in set  $k$ . Other notations are similar.

$R^{(1)}$  and  $R_k^{(2)}$  are two random matrices sampled from uniform distribution ranging

from [0, 1]. N is the number of feature map sets, and t indicates the degree of overlap of the neurons. The final mask is the union of  $M_k^{(1)}$  and  $M_k^{(2)}$  :

$$M_k = M_k^{(1)} \vee M_k^{(2)} \quad (3)$$

Only neurons at the place where the element values of  $M_k$  are 1 can pass the mask.

### 3 Experiment and Discussion

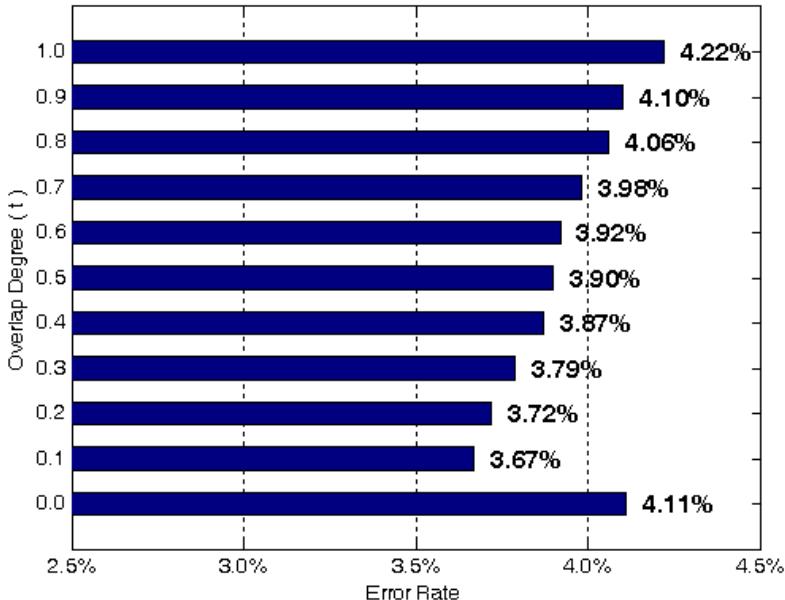
We test our method on CASIA-HWDB1.1 database using cuda-convnet framework on NVIDIA GeForce GTX 690 GPU. CASIA-HWDB1.1 is an offline dataset of isolated characters which consists of 3,755 GB2312-80 level-1 Chinese characters and 171 alphanumeric and symbols written by 300 writers. More details can be found in [17]. A subset of CASIA-HWDB1.1 containing the first 1000 classes of Chinese characters is used in our experiment. Each class has about 240 images for training and about 60 images for testing. The number of total training sample is 239,124 and the number of total testing samples is 59660. Cuda-convnet [18] is an open source feed-forward neural network library implemented in C++/CUDA. It is extremely fast so that we can train each of our networks in less than 24 hours.

The architecture of CNN we adopt consists of 4 convolutional layers, each of which is followed by an element-wise max-pooling layer and a max-pooling layer. In each convolutional layer, both 3×3-size and 5×5-size local receptive fields are used to produce two sets of feature maps. The stride of all local receptive fields is 1. The rectified linear units are adopted as the activation function. The element-wise max-pooling layer compares the elements of that two feature map sets and chooses the maximum. The window size and the window stride of all max-pooling layers are 3 and 2, respectively, so that the feature maps are down sampled by a factor of 2. In all convolutional layers and max-pooling layers, padding is used when necessary. The last two layers are the full connected layer. The final output is obtained by soft max regression which denotes the probability distribution of the input correspond to the class labels. For description convenience, we denote convolutional layer as Conv, element-wise max-pooling layer as EleMax, max-pooling layer as MaxP and full-connected layer with N neurons as Full\_N. We also further denote the convolutional layer by Conv\_A\_B[\_C] where B and C are the sizes of local receptive fields, A is the number of feature maps correspond to each local receptive field. The bracket means that C is optional. In this notation strategy, the CNN architecture we adopt can be described as Conv\_32\_3\*3\_5\*5 – EleMax – MaxP – Conv\_64\_3\*3\_5\*5 - EleMax – MaxP –Conv\_128\_3\*3\_5\*5 - EleMax – MaxP –Conv\_128\_3\*3\_5\*5 - EleMax – MaxP – Full\_1024 – Full\_1000.

In the training process, each sample is randomly distorted by the method described in [19]. All weights of the network are initialized by randomly sampling from  $(\mu, \sigma) = (0, 0.01)$ , and the initial value of all biases is 0.5. All parameters are updated by stochastic gradient descent with a min batch of 128, a momentum of 0.9 and a weight decay of 0.001 for full-connected layers. The network is trained in 120 loops. The learning rate is 0.001 for the first 100 loops and multiplied by 0.1 for the last 20 loops.

### 3.1 Select Parameter t

Compared to the traditional CNN model, our model introduces only one additional parameter  $t$ , which indicates the degree of overlap of neurons in the same position. To get the optimal value, we assign  $t$  with values uniformly sampled from  $[0, 1]$  with a stride of 0.1 and train a CNN model with each  $t$ . The result is shown in Fig.2, from which we can see that the optimal value of  $t$  is 0.1 and best performance is an error rate of 3.67%. We use this error rate as a performance of our model and compare it with the performance of other models in the following parts.



**Fig. 2.** The performance of CNN with different values of  $t$ .

It is worth mentioning that even when no competition strategy is used in training process ( $t=0$ , which means that neurons of the modified convolutional layer are trained totally randomly) the network performs rather well (error rate is 4.11%). One possible reason is that the use of the conventional max-pooling operation and the rectified linear units makes the model tend to prefer neurons with large activation values. Furthermore, the network architecture itself inherently provides some frequency selection and translation invariances, as described in [20].

### 3.2 Compare to Traditional CNNs

We first check our model's ability of selecting the optimal local receptive field by enumerating possible combinations of local receptive fields of the convolutional layers. The number of all local receptive field combinations is 16 (4 convolutional layers and 2 local receptive field candidates for each convolutional layer). To reduce

the computational load, we select 5 of them under the principle that the size of local receptive field of a convolutional layer cannot be larger than that of the previous convolutional layers. It is reasonable since the input feature maps of a convolutional layer are not larger than that of the previous convolutional layers. The structures of all comparison models can be described as Conv\_32\_a\*a – MaxP – Conv\_64\_b\*b – MaxP – Conv\_128\_c\*c – MaxP – Conv\_128\_d\*d – MaxP – Full\_1024 – Full\_1000, where  $a, b, c, d \in \{3, 5\}$ . Table 1 shows the performance of CNNs with different local receptive field combinations. The error rate of the optimal model is 4.10%, which is higher than that of our model by 0.43%. This demonstrates that our model is able to select optimal local receptive field by itself so that the performance is improved.

**Table 1.** CNN performance with different local receptive field combinations

Model	a	b	c	d	Error Rate
1	3	3	3	3	4.37%
2	5	3	3	3	4.22%
3	5	5	3	3	4.18%
4	5	5	5	3	4.10%
5	5	5	5	5	4.16%

**Table 2.** The performance of different CNN models

Model	Parameter Number	Error Rate
Traditional CNN	3.70 Million	3.90%
Our model	3.11 Million	3.67%
Dropout [12]	3.70 Million	3.45%
Maxout[15]	3.65 Million	3.73%
Our model + Dropout	3.11 Million	3.36%

Based on the structure of the optimal model in Table 1, we train a new CNN model with more feature maps in each convolutional layer. The structure of the new model is Conv\_48\_5\*5 – MaxP – Conv\_96\_5\*5 – MaxP – Conv\_192\_5\*5 – MaxP – Conv\_192\_3\*3 – MaxP – Full\_1024 – Full\_1000. Our goal is to train a CNN with the same number of parameters as our model so that we can tell whether the improvement of the performance is due to the refinement of network structure or due to the increased number of weights. Experimental result (see Table 2) shows that our model performs better than traditional CNN even with less parameters, which proves that it is the new structure makes our model perform better.

By randomly blocking half of the neurons in the first full connected layer, we train two dropout networks. One is the traditional dropout network, and the other is based on our model. We also train a maxout network by grouping the neurons of the first full connected layer into 512 groups (2 neurons each group) and replace the non-linear function by maxout nonlinear form. We do not use the maxout activation units in the convolutional layers simply because its performance is really not very good in our experiments, compared to the other models. Table 2 shows the performance of each model. Dropout as a successful regularization method improves the performance

of the traditional CNN model by 0.45%, while our method improves the performance by 0.23%. Dropout dose perform slightly better than our method, but when we add the dropout strategy into our model, we achieve an error rate of 3.36%, which is better than both two models.

We notice that although we apply the exactly same dropout operation to the traditional CNN model and our model, the performance improvements are different. The dropout operation improves the performance of the traditional CNN model by 0.45%, which is slightly larger than the magnitude of the performance improvement of our model (0.31%). We believe that it is because our model introduces some disturbances during training procedure, which break some interactions between large numbers of units [21], thus the network is easy to train by the current standard first order gradient descent. On the contrary, the traditional CNN model makes no attempts to break the neuron interactions. So it is trained less fully than our model under the same training process setup. As a result, the dropout method affects our model slightly less than the traditional CNNs.

## 4 Conclusion and Future Work

This paper presents a novel method to solve the problem of selecting local receptive fields of the convolutional layers in traditional CNNs. Instead of selecting local receptive field according to some prior information, we simply produce several sets of feature maps with different local receptive fields and let the model itself to choose the best neurons. A new element-wise max pooling layer is introduced to do this job. To ensure that all neurons in the model are fully trained, a random mask is generated for every feature map of the inputs of the element-wise max-pooling layers. Experiments on handwriting Chinese character recognition illustrate the excellent performance of our method.

At first glance, the success of our model comes from its ability of adaptively selecting the appropriate local receptive field of the convolutional layer. But from the feature selection perspective, our method can be also regarded as a method of selecting the right features from a feature bank produced by different local receptive fields. We believe that other types of features can be well incorporated in this framework, and it is exactly what we will do in our future work.

## References

1. Wu, C., Fan, W., He, Y., Sun, J., Naoi, S.: Cascaded Heterogeneous Convolutional Neural Networks for Handwritten Digit Recognition. In: 21st IEEE International Conference on Pattern Recognition, pp. 657–660. IEEE Press, Tsukuba (2012)
2. Ciresan, D., Schmidhuber, J.: Multi-Column Deep Neural Networks for Offline Handwritten Chinese Character Classification. Technical report, IDSIA (2013)
3. Yaniv, T., Ming, Y., Marc'Aurelio, R., Lior, W.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: 27st IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, Columbus (2014)

4. Alexander, T., Christian, S.: DeepPose: Human Pose Estimation via Deep Neural Networks. In: 27th IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, Columbus (2014)
5. Alex, K., Ilya, S., Geoffrey, H.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems 25. NIPS Foundation, Nevada (2012)
6. Ossama, A., Li, D., Dong, Y.: Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition. In: Interspeech 2013, ISCA (2013)
7. Yann, D., Yoshua, B.: Big Neural Networks Waste Capacity. In: International Conference on Learning Representations, Scottsdale (2013)
8. Coates, A., Ng, A., Lee, H.: An Analysis of Single-layer Networks in Unsupervised Feature Learning. In: 14th International Conference on Artificial Intelligence and Statistics, Reykjavik, pp. 215–223 (2011)
9. Coates, A., Ng, A.: Selecting Receptive Fields in Deep Networks. In: Advances in Neural Information Processing Systems 24. NIPS Foundation, Granada (2011)
10. Jia, Y., Huang, C., Darrell, T.: Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features. In: 25th IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press (2012)
11. Kong, S., Jiang, Z., Yang, Q.: Collaborative Receptive Field Learning. arXiv Preprint arXiv:1402.0170 (2014)
12. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. arXiv preprint arXiv:1207.0580 (2012)
13. Wan, L., Zeiler, M., Zhang, S., Cun, Y.L., Fergus, R.: Regularization of Neural Networks using Dropconnect. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1058–1066 (2013)
14. Zeiler, M.D., Fergus, R.: Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. In: International Conference on Learning Representations, Scottsdale (2013)
15. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout Networks. In: International Conference on Learning Representations, Scottsdale (2013)
16. Wu, C., Fan, W., He, Y., Sun, J., Naoi, S.: Handwritten Character Recognition by Alternately Trained Relaxation Convolutional Neural Network. Submitted to ICFHR (2014)
17. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: CASIA Online and Offline Chinese Handwriting Databases. In: 2011 International Conference on Document Analysis and Recognition, pp. 37–41. IEEE Press (2011)
18. cuda-convnet project, <https://code.google.com/p/cuda-convnet/>
19. Simard, P., Steinkraus, D., Platt, J.C.: Best Practice for Convolutional Neural Networks Applied to Visual Document Analysis. In: 2003 International Conference on Document Analysis and Recognition. IEEE Press (2003)
20. Saxe, A., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: On Random Weights and Unsupervised Feature Learning. In: Proceedings of the 28th International Conference on Machine Learning, pp. 1089–1096 (2010)
21. Bengio, Y.: Deep learning of representations: Looking forward. In: Dedić, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) SLSP 2013. LNCS, vol. 7978, pp. 1–37. Springer, Heidelberg (2013)

# Character Segmentation for Classical Mongolian Words in Historical Documents

Xiangdong Su, Guanglai Gao, Weihua Wang, Feilong Bao, and Hongxi Wei

School of Computer Science, Inner Mongolia University  
Hohhot, China 010021  
csggl@imu.edu.cn

**Abstract.** There are many classical Mongolian historical documents which are reserved in image form, and as a result it is inconvenient for us to search and mining the desired content. In order to facilitate the word recognition in the document digitization procedure, this paper proposes a novel approach to segment the historical words in which the characters are intrinsically connected together and possess remarkable overlapping and variation. The approach consist of three steps: (1)significant contour point (SCP) detection on the approximated polygon of the word's external contour, (2)baseline locating based on the logistic regression model and (3)segment path generation and validation based on the heuristic rules and the neural network. The SCP helps in the baseline locating and segment path generation. Experiment on the historical Mongolian Kanjur demonstrates that our approach could effectively locate the words' baselines and segment the words into characters.

**Keywords:** Classical Mongolian, Character Segmentation, Logistic Regression, Heuristic Rule, Neural Network.

## 1 Introduction

Historical documents are precious culture heritages of human beings. At present, there is a growing trend towards digitization of them to make these significant documents easier to be preserved, accessed, and shared. In Inner Mongolia Autonomous Region of China, there are a large number of ancient Mongolian books. One of the most famous books is the Mongolian Kanjur. It involves history, literature, religion, sociology and many other aspects, and is considered to be an encyclopedia. Although it has been converted into the digital image collection, the more efficient and effective way to keep it while making it publicly available and easy to be searched is to convert its image collection into text by means of optical character recognition (OCR).

The varieties of word recognition techniques in OCR fall into two classes: holistic recognition and segment-based recognition, depending on whether segmentation is required. Although holistic recognition methods free us from the segmentation and usually outperform the latter, their complexities grow as the vocabulary gets larger [1]. They uses the lexicon as knowledge-base and are suitable for small and static lexicon based application [2]. Their recognition accuracies is generally linear to the

size of the lexicon [3]. On a conservative estimate, the number of all the lexicon items in the Mongolian Kanjur is greater than 100,000, which is a very large number. For holistic recognition, training such a classifier, the number of whose output labels is equal to the size of the lexicon, is certainly difficult and time-consuming. On the contrary, as to segmentation-based methods, the number of the classes of the recognition unit, here character, is much smaller than the size of the lexicon. Using such methods could decrease the recognition difficulty dramatically. Thus, the segmentation-based method is more practical than the holistic method in classical Mongolian recognition.

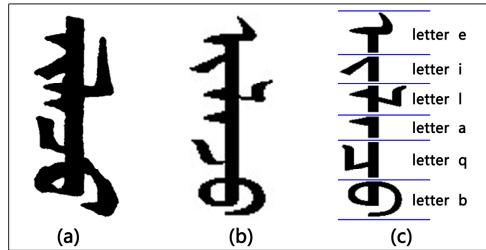
In fact, character segmentation has long persisted as long as word recognition. The segmentation performance directly affects the recognition accuracy [4]. A word is unlikely to be recognized correctly when it was segmented improperly, because the segmentation errors bring negative impact to the character recognition. Conventional approaches take two steps to segment each machine-printed Mongolian word with the horizontal lines whose widths are equal to the word's width [5]. At first, the baseline is determined in terms of the word's vertical projection. Then the specified rows are selected as the segmentation lines when the constraint is satisfied. However, such methods are ineligible for segmenting the classical Mongolian words in the Mongolian Kanjur. The problems mainly concentrate on two aspects. First, only relying on the vertical projections, they are failed to locate the baselines which are crucial for the segmentation line validation. Second, due to character overlapping and deformation, it is intractable to find the horizontal word-width segmentation lines which split each pair of successive characters properly. Therefore, this paper put forward a novel approach to segment the classical Mongolian words so as to facilitate the word recognition in the digitization of the Mongolian Kanjur.

To facilitate understanding of the challenges in character segmentation of classical Mongolian words, it is necessary to introduce the Mongolian Kanjur. The Mongolian Kanjur is written in classical Mongolian which is an agglutinative language. The characters consisted of a word are joined together along the baseline. That is, the characters are intrinsically touched. The Mongolian Kanjur is made by woodblock printing in the period of Qing Dynasty 300 years ago. This ancient printing technology leads to a situation that multiple artisans created many variants for the same word. More importantly, character overlapping is a common phenomenon. Even some characters overlap with non-adjacent characters. Fig. 1 presents a word in the Mongolian Kanjur including its machine-printed form and its characters. In addition, ink spreading causes the deformation of words and the spur noise in the words.

Considering that the characters in classical Mongolian words are connected along the baseline, the proposed approach takes completely different strategy to segment these words. It does not try to find a word-width horizontal line to segment each pair of neighboring characters rather than only segment the place where they touched. The segment path is straight but not necessarily horizontal. The workflow of the approach can be summarized as follows. The first step is detecting the SCPs on the approximated polygon of each word's external contour. The second step is locating the baseline with the logistic regression model. The final step generates the candidate segment paths (CSPs) based on the SCPs with the heuristic rules and validates them

with the neural network. The experimental performance achieves 89.84% correct segmentation rate at character level and 79.66% completely correct rate at word level, which demonstrates our approach is applicable and effective.

The remainder of this paper is organized as follows. Section 2 reviews some related works in character segmentation. Section 3 details the proposed approach used in character segmentation of classical Mongolian words. The experiment is demonstrated in section 4. The conclusion is presented in section 5.



**Fig. 1.** (a)The woodblock printed “𠂇”，(b) its machine-printed form and (c) its characters

## 2 Related Works

Character segmentation is the intensive research in word recognition, and many methods have been proposed in the published literature. A substantial part of these methods belongs to the over-segmentation scheme, in which the CSPs are generated in the preliminary dissection stage, and the unnecessary ones are rejected in the validation stage. Finding the CSPs generally involves the histogram, contour, skeleton and region information of the words image. For each CSP, a decision whether it will be kept or not have to be made in the validation stage. Verma in [6] used a heuristic segmentation to segment each word and trained a neural network to validate the segmentation paths with the contour code feature. Lei et al. in [7] described a segmentation system for touching handwritten numeral strings in which candidate segmentation points was generated form the corner points and horizontal projection. In [8], Viterbi algorithm and background thinning method was applied to find the CSPs in non-touching case. The redundant CSPs were eliminated by heuristics. Background and foreground information was applied to find the CSPs in touching case. The optimal CSPs were selected with the mixture probabilistic density function. Vellasquesa et al. in [9] employed a classifier to identify the unnecessary segmentation cut to reduce the computation cost and increase the overall performance.

The following works are relevant to Mongolian segmentation. Gao et al. in [10] investigated the peculiarities of classical Mongolian historical documents and designed a segmentation-based approach to recognize the words in them. This method used word-width horizontal line to split the neighboring characters. It could not solve the serious overlapped cases and resulted many under-segmented fragments. Peng et al. in [11] investigated multi-font machine-printed Mongolian document recognition. During character segmentation, segmentation points were identified by analyzing the properties of projection profiles and connected components.

### 3 Segmentation Approach

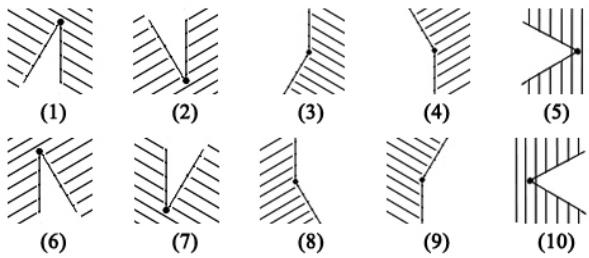
As described above, character variation is remarkable and character overlapping is quite frequent in the Mongolian Kanjur. In case of character overlapping, it is intractable to find a word-width horizontal line to split the overlapped neighboring characters. Even for the words without character overlapping, it is more difficult to determine the segmentation paths of the historical words than their machine-printed forms because the baselines are hard to locate. In such scenario, an applicable strategy is directly cutting the connected places and then using the connected component algorithm to obtain the isolate characters. Our approach adopts this strategy and takes the following three steps to segment the historical words.

#### 3.1 Significant Contour Point Detection

In classical Mongolian words, the connected places between character strokes and baseline, or the touched places between the neighboring characters form some corner points. They are named as significant contour points (SCPs) in this paper. These points, especially the points on the external contour, play important roles in character segmentation. In one hand, they provide the critical clues for baseline determination. In another hand, they can be used to derive the entry point (EP) and the exit point (EXP) of the CSP between two successive characters with some heuristic rules.

Since each SCP represents a point at which the contour direction changes, they can be detected on the approximated polygon of word's external contour. It can simplify the detection procedure, restrain disturbance and false tripping caused by noises on the contour, and further achieve a certain degree of robustness. This paper takes Ramer–Douglas–Peucker algorithm [12,13] to approximate the external contour to a polygon. The fit criterion parameter is set to 10 experimentally. The output polygon is fed into a detector to find the desired SCPs.

In this paper, totally ten kinds of the SCP are detected (See Fig. 2). The former five kinds (1)-(5) correspond to the SCPs on the left contour, and the remnant five kinds (6)-(10) correspond to the SCPs on the right contour. Here, the left contour represents the part from the lowest point to the highest one in clockwise direction; the right contour represents the part from the highest point to the lowest one in the same direction. The SCP in (1) is a vertex on the polygon, one of whose adjacent edge is nearly vertical and the other adjacent edge is on the left of the former. At the same time, the intersection angle should be upward and smaller than  $90^\circ$ . The shade represents black pixels of the word image. The others are also straightforward. An example is shown in Fig. 3(d). For some words, it is possible to detect several corner points in a small region due to local zigzag noises on their contours. Keeping one of them is sufficient for character segmentation. In such a case, the central one is reserved as a SCP according to a preset radius of neighborhood.



**Fig. 2.** Ten Types of the SCPs

### 3.2 Baseline Locating

The second step is responsible for locating the word's baseline. This is equivalent to fix the two columns of its left and right boundaries in the word image. For machine-printed Mongolian word, the baseline is vertical and its width is unique. Commonly, each boundary can be determined by search the column where a steep change occurs in the vertical projection of the word image. However, such method is failed for the woodblock printed classical Mongolian words, especially the words with fewer characters. In one hand, the baseline is not completely vertical. In another hand, there may be no steep changes near the baseline's boundaries. This is mainly caused by the word slant, the variation of stroke thickness and ink spreading. Therefore, a more sophisticated method is required to estimate the baseline.

According to the above description, the SCPs always appear near the baseline. This is determined by the way of their formation. Therefore, it is intuitive to incorporate the SCPs to increase the accuracy of baseline locating. In this paper, multiple feature variables are fed into the logistic model to predict the possibility of each point on the x-axis as the boundary of the baseline. The possibility of  $i$ th point is

$$p(i) = \frac{1}{1 + e^{-(\mathbf{W} \cdot \mathbf{F})}} \quad (1)$$

where  $\mathbf{W}$  is the weight parameter vector, and  $\mathbf{F}$  is the vector of the feature variables. The feature variables include three types information: scale and coordinate information, projection information and the morphological information (SCPs). They are all presented in the Table 1. Here  $w$  is the width of the image;  $h$  is the height of the image;  $i$  is the x-coordinate (column number);  $p_i$  is the height of the  $i$ th column in the histogram;  $p_h$  is the height of the highest column in the histogram;  $s$  is the number of SCPs in the image;  $sl5$  is the number of SCPs between  $i-5$ th column and  $i$ th column in the word image;  $sr5$  the number of SCPs between  $i$ th column and  $i+5$ th column;  $sl10$  and  $sr10$  are defined similar.

Training data is generated from the training words whose baselines are manually annotated. When the possibilities are ready, each column is classified into 0 or 1 according to whether its possibility is bigger than the threshold. For the 1's intervals, the top 2 widest intervals are selected, and their center positions are used as the left and

right boundaries of the baseline. Provide there are less than two internals, the threshold is adjusted. The baseline's width needs to range from 35 to 70.

**Table 1.** Feature variables used in the baseline locating

<b>F1:</b> $w$	<b>F2:</b> $h$	<b>F3:</b> $w/h$	<b>F4:</b> $i/w$	<b>F5:</b> $(w-i)/w$
<b>F6:</b> $p_i$	<b>F7:</b> $p_j/h$	<b>F8:</b> $p_i/p_h$	<b>F9:</b> $p_i/p_{i-5}$	<b>F10:</b> $p_i/p_{i+5}$
<b>F11:</b> $p_i/p_{i-10}$	<b>F12:</b> $p_i/p_{i+10}$	<b>F13:</b> $s$	<b>F14:</b> $s/5$	<b>F15:</b> $s/5/s$
<b>F16:</b> $sr5$	<b>F17:</b> $sr5/s$	<b>F18:</b> $sl10$	<b>F19:</b> $sl10/s$	<b>F20:</b> $sr10$
<b>F21:</b> $sr10/s$				

### 3.3 Segment Path Generation and Validation

This step deals with the generation and validation of the CSPs. The core idea of the proposed approach is directly generating the segmentation path to split the connected place between each pair of the characters. The segmentation paths are straight but not necessary horizontal. This amounts to determine the EP and the EXP of the segmentation path. This paper takes two steps to achieve the goal. At first, the EPs are obtained with several heuristic rules form the SCPs. For each EP, four interrelated EXPs are considered. That is, each EP gives birth to four CSPs. Secondly, a back propagation (BP)-based neural network is employed to validate the CSPs. The unnecessary segmentation paths are eliminated in this step. The SCPs are processed sequentially with the priority rules to generate the EPs. The rules are listed as follows:

- **Rule 1.** If the SCP  $P_i$  belongs to type (1), its neighboring SCP  $P_j$  belongs to type (3) and there is no convex between them, then the  $P_j$  is treated as an EP and  $P_i$  is neglected. Similarly, if the SCP  $P_i$  belongs to type (2), its neighboring SCP  $P_j$  belongs to type (4) and there is no convex between them, then  $P_j$  is treated as an EP and  $P_i$  is neglected. The neglected SCPs will not be reconsidered.
- **Rule 2.** If the SCP  $P_i$  and its neighboring SCP  $P_j$  belong to type (1) and there is no convex between them, then the point which is on the left contour and in the middle of them in vertical direction is treated as an EP. Similarly, if the SCP  $P_i$  and its neighboring SCP  $P_j$  belong to type (3) and there is no convex between them, then the point which is on the left contour and in the middle of them in vertical direction is treated as an EP.
- **Rule 3.** Each SCP  $P_i$  belonging to type (5) is treated as an EP.
- **Rule 4.** If there is a convex between a pair of neighboring SCPs on the left contour, these two SCPs will be treated as EPs at the same time.
- **Rule 5.** After the above-mentioned rules are handled, the rest unprocessed SCPs on the left contour are wholly treated as EPs.
- **Rule 6.** After all the SCPs on the left contour are processed, if the vertical distance between a pair of neighboring EPs is higher than the highest the character, the point on the left contour who is between them in vertical direction and its y-coordinate is equal to the y-coordinate of an SCP on the right contour is treated as an EP.

For each EP  $p(x, y)$ , the four interrelated EXPs are considered: (1) the nearest SCP on the right contour whose y-coordinate is smaller than  $y$ , (2) the nearest SCP on the right contour whose y-coordinate is larger than  $y$ , (3) the nearest point on the right contour and (4) the point on the right contour whose y-coordinate is equals to  $y$  and its distance to the right boundary of the baseline is shortest.

The BP neural network used in CSP validation are fully connected and have four layers: one input layer, two hidden layers and one output layers. The number of the input layer's neurons equals the dimension of the input feature vector and the number of the output layer's neurons is 1. The numbers of the two hidden layers' neurons are set experimentally. The training words are the same as the ones used in baseline locating. Their CSPs are manually classified in the training set.

The input feature vector of the neural network is built by integrating the properties of the baseline and the CSPs. It includes 36 feature variables which reflect the location and size relationships between these properties. Table 2 describes all the variables. In the table,  $nb$  is the number of the black point on the CSP;  $len$  is the length of the CSP;  $bw$  is the width of the baseline;  $lbr$  and  $rbr$  are the left boundary and right boundary of the baseline;  $x_1$  is the x-coordinate of the EP;  $x_2$  is the x-coordinate of the EXP;  $\theta$  is the angle between the CSP and x-axis;  $IsScp(EXP)$  judge whether the EXP is an SCP. Fig. 3 demonstrates the character segmentation for the word “**жанр**”.

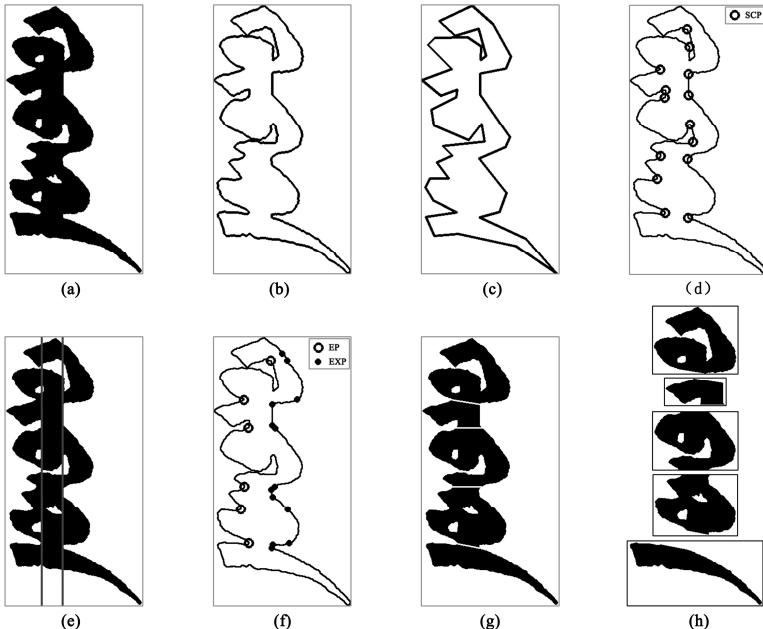
**Table 2.** Feature variables used in the baseline locating

<b>F1:</b> $nb$	<b>F2:</b> $len$	<b>F3:</b> $nb/len$	<b>F4:</b> $nb-len$
<b>F5:</b> $(nb-len)/len$	<b>F6:</b> $bw$	<b>F7:</b> $nb/bw$	<b>F8:</b> $nb-bw$
<b>F9:</b> $(nb-bw)/bw$	<b>F10:</b> $x_2-x_1$	<b>F11:</b> $nb/(x_2-x_1)$	<b>F12:</b> $nb-(x_2-x_1)$
<b>F13:</b> $(nb-(x_2-x_1))/(x_2-x_1)$	<b>F14:</b> $len/bw$	<b>F15:</b> $len-bw$	<b>F16:</b> $(len-bw)/bw$
<b>F17:</b> $len/(x_2-x_1)$	<b>F18:</b> $len-(x_2-x_1)$	<b>F19:</b> $(len-(x_2+x_1))/(x_2-x_1)$	<b>F20:</b> $bw/(x_2-x_1)$
<b>F21:</b> $bw-(x_2-x_1)$	<b>F22:</b> $(bw-(x_2-x_1))/(x_2-x_1)$	<b>F23:</b> $x_1$	<b>F24:</b> $lbx$
<b>F25:</b> $x_1-lbx$	<b>F26:</b> $(x_1-lbx)/len$	<b>F27:</b> $(x_1-lbx)/bw$	<b>F28:</b> $(x_1-lbx)/(x_2-x_1)$
<b>F29:</b> $x_2$	<b>F30:</b> $rbx$	<b>F31:</b> $x_2-rbx$	<b>F32:</b> $(x_2-rbx)/len$
<b>F33:</b> $(x_2-rbx)/bw$	<b>F34:</b> $(x_2-rbx)/(x_2-x_1)$	<b>F35:</b> $\cos\theta$	<b>F36:</b> $IsScp(EXP)$

## 4 Experiments

The experiment is conducted on 9000 words which are extracted from the Mongolian Kanjur. 4000 words are used for training. The rest 5000 words are processed with the above-mentioned approach to check the effectiveness of the proposed approach in the test experiment. It is required to inspect the segments to evaluate the performance. To minimize the bias to the segmentation result, a blind inspection can be performed. In the inspection, some people, who are unrelated to the research, categorize each segment into one of the five kinds: correct segmentation, over-segmentation, under-segmentation, failed segmentation and bad segmentation. The correct segmentation denotes that the segment is a character. This means that the segmentation path separate

the neighboring characters properly. The over-segmentation is defined as that a character is cut into two segments by a redundant segmentation path. The under-segmentation denotes that the segment is a pair of undivided neighboring characters. The failed segmentation represents a situation that the resulting segment includes more than two connected characters which have not been separated. This is caused by two or more neighboring segmentation paths are missed. The bad segmentation represents the rest of inappropriate segments that do not belong to under-segmentation, over-segmentation, or failed segmentation.



**Fig. 3.** Character segmentation for the classical Mongolian word “жагаан”. **a** the word, **b** its external contour, **c** the approximated polygon of the external contour, **d** the SCPs on the external contour, **e** the baseline, **f** the EPs and EXPs, **g** the segmentation result, **h** the segments.

The final evaluation metrics group into character level and word level. At character level, five evaluation metrics are calculated by dividing the numbers of the segments in the above five categories by the total number of the characters consisting of the test words respectively. At word level, completely correct rate and accepted correct rate are computed. All the metrics are listed as follows.

$$\text{correct segmentation rate} = \frac{\#\text{correct segmentation}}{\#\text{characters}} \quad (2)$$

$$\text{under-segmentation rate} = \frac{\#\text{under-segmentation}}{\#\text{characters}} \quad (3)$$

$$\text{over-segmentation rate} = \frac{\#\text{over-segmentation}}{\#\text{characters}} \quad (4)$$

$$\text{failed segmentation rate} = \frac{\# \text{ failed - segmentation}}{\# \text{ characters}} \quad (5)$$

$$\text{bad segmentation rate} = \frac{\# \text{ bad - segmentation}}{\# \text{ characters}} \quad (6)$$

$$\text{completely correct rate} = \frac{\# \text{ words whose characters are completely correctly segmented}}{\# \text{ words}} \quad (7)$$

$$\text{accepted correct rate} = \frac{\# \text{ words without failed - segmentation and bad - segmentation}}{\# \text{ words}} \quad (8)$$

Except the correct segmentation, the other segmentations totally belong to error segmentation. However, for simple errors (over-segmentation and under-segmentation), most of them can be corrected by using the feedback from the recognition stage. So the accepted correct rate is calculated at word level. The evaluation results are shown in Table 3 and Table 4.

The performance demonstrates that the proposed approach could efficiently segment the classical Mongolian words in the historical documents. This is due to the utilization of the SCPs. The logistic regression model could predict the baseline more accurately than the method only relying on the vertical projection. This ensures the success of segmentation path validation which relies heavily on the baseline. The constraint of the proposed approach is that it cannot deal with multiple-touched characters, although such scenarios are very rare. The main reasons leading to segmentation errors are two fold: (1) failure to detect the baseline leads to the failure in the validation stage and (2) multi-touching case results some under-segmentation errors. The first scenario mainly arises when the prefetched SCPs are few and word distortion is remarkable.

**Table 3.** performance (%) at character level

correct segmentation rate	under-segmentation rate	over-segmentation rate	failed segmentation rate	bad segmentation rate
89.84	3.18	1.35	0.21	1.72

**Table 4.** performance (%) at word level

completely correct rate	accepted correct rate
79.66	91.04

## 5 Conclusions

This paper investigates the character segmentation for classical Mongolian words in woodblock-printed historical documents. The challenges are discussed firstly. Emphasis is placed on the baseline locating and segmentation path generation. Morphological information on the external contour (SCP) is used in the former

procedures. The performance achieves 89.84% correct segmentation rate at character level and 79.66% completely correct rate at word level, which looks satisfactory considering the fact that the characters in the historical words are intrinsically touched together with remarkable variation and overlapping.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China (Grant No. 61263037) and program of higher-level talents of Inner Mongolia University (SPH-IMU).

## References

1. Zand, M., Nilchi, A.N., Monadjemi, S.A.: Recognition-based Segmentation in Persian Character Recognition. World Academy of Science, Engineering and Technology 2, 162–166 (2008)
2. Saba, T., Rehman, A., Elarbi-Boudihir, M.: Methods and Strategies on Off-line Cursive Touched Characters Segmentation: a Directional Review. Artificial Intelligence Review (2011)
3. Verma, B., Lee, H.: Segment Confidence-based Binary Segmentation (SCBS) for Cursive Handwritten Words. Expert Systems with Applications 38, 11167–11175 (2011)
4. Lee, H., Verma, B.: Binary Segmentation Algorithm for English Cursive Handwriting Recognition. Pattern Recognition 45, 1306–1317 (2012)
5. Li, W., Gao, G., Hou, H., Li, Z.: A Design and Implementation of Element Segmentation in the Recognition of Printed Mongolian Characters. Inner Mongolia University 34, 357–360 (2003)
6. Verma, B.: A Contour Code Feature Based Segmentation For Handwriting Recognition. In: Proceedings of ICDAR, vol. 2. IEEE Computer Society (2003)
7. Lei, Y., Liu, C.S., Ding, X.Q., Fu, Q.: A Recognition Based System for Segmentation of Touching Handwritten Numeral Strings. In: Liu, C.S., Ding, X.Q., Qiang, F. (eds.), pp. 294–299 (2004)
8. Liang, Z., Shi, P.: A Metasynthetic Approach for Segmenting Handwritten Chinese Character Strings. Pattern Recognition Letters 26, 1498–1511 (2005)
9. Vellasquesa, E., Oliveiraaa, L.S., Britto Jr., A.S., Koericha, A.L.: Filtering Segmentation Cuts for Digit String Recognition. Pattern Recognition 41, 3044–3053 (2008)
10. Gao, G., Su, X., Wei, H., Gong, Y.: Classical Mongolian Words Recognition in Historical Document. In: Proceedings of ICDAR, pp. 692–697. IEEE Computer Society (2011)
11. Peng, L., Liu, C., Ding, X., Jin, J., Wu, Y., Wang, H., Bao, Y.: Multi-font Printed Mongolian Document Recognition System. International Journal of Document Analysis Recognition 13, 93–106 (2010)
12. Ramer, U.: An Iterative Procedure for the Polygonal Approximation of Plane Curves. Computer Graphics and Image Processing 1, 244–256 (1972)
13. Douglas, D., Peucker, T.: Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature. Cartographica: The International Journal for Geographic Information and Geovisualization 10, 112–122 (1973)

# MCDF Based On-Line Handwritten Character Recognition for Total Uyghur Character Forms

Askar Hamdulla, Wujiahemaiti Simayi, Mayire Ibrayim, and Dilmurat Tursun

Xinjiang University, Urumqi, 830046, China  
askar@xju.edu.cn

**Abstract.** This paper proposed the Modified Center Distance Feature (MCDF) and its different forms for Uyghur handwritten character recognition. By combination with some low dimensional features, MCDF gifted remarkable recognition accuracy of 87.6% for total Uyghur character forms. This result is higher than previous record by more than 11 points. Samples from 400 volunteers are used in experiments.

**Keywords:** Uyghur characters, On-line handwritten recognition, Low dimensional features, Modified center distance feature.

## 1 Introduction

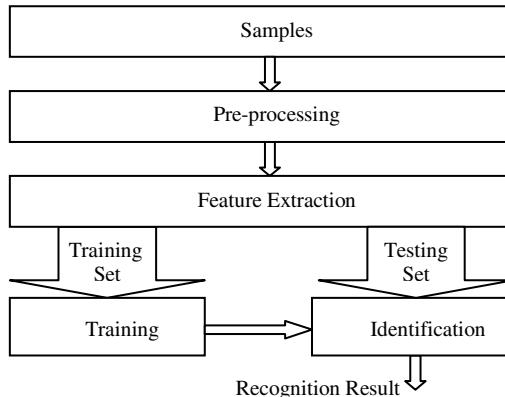
Character recognition technology is one of the most important research fields in pattern recognition [1]. Handwritten Character recognition refers the technology that handwritten characters can be recognized by computer or intelligent system, and the templates corresponding to the characters are identified. Handwritten character recognition technology has two main sub-fields, such as online and off-line handwritten character recognition [2]. Online character recognition recognizes character patterns captured from a pen-based or touch-based input device where trajectories of pen-tip or finger-tip movements are recorded, while offline recognition recognizes character patterns captured from a scanner or a camera device as two dimensional images [3].

Uyghur is one of main languages in Altaic language system. Uyghur texts are made from 32 basic characters [4]. With at least two and at most eight forms, Uyghur has total of 128 character forms. Although there are some regulations to write characters correctly, various kinds of ways cannot be prevented in actual handwriting. Writing style and order of handwritten characters are influenced by different writers and the writer's mood, writing context, etc.

## 2 On-line Uyghur Handwritten Character Recognition System

The function units of the handwritten character recognition system applied for Uyghur characters in this paper is shown in Fig 1. The first-hand data collected from different

writers are inputted to the system at first. The pre-processing unit mainly deals with the noise elimination and normalization on the raw data [5]. One of the most critical works is processed during the feature extraction. The data of the features extracted from the characters are classified to several sub groups. The template library is built using quick search classification method [6]. The recognition results are provided by minimum distance classification using Euclidean distance.



**Fig. 1.** Block diagram of on-line Uyghur handwritten character recognition system

### 3 Feature Extraction

Feature extraction is one of the most challenging and innovative task in handwritten recognition [7]. Effective feature observation and extraction is crucial to improve the accuracy and efficiency of character recognition system. This section gives a brief introduction for the features used in experiments and their feature extraction methods.

#### 3.1 Stroke Number Feature and Additional Part's Location Feature

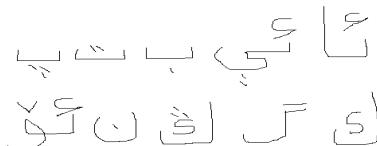
Stroke number feature refers to the number of strokes that form the handwritten character trajectory. The additional strokes besides the main stroke are called the additional part of character, while the main stroke is named as main part. The additional part is located over or within or under the main part.

#### 3.2 Bottom-Up (BUDR) and Left-Right (LRDR) Density Ratio

If the normalized sample is observed with  $5 \times 5$  square grids, we easily calculate the ratio between the pixels of the written letter in the first two rows and in the last two rows or between the first two columns and the last two columns. The first case is called bottom-up ratio while the second one is called left-right ratio [9].

### 3.3 Shape Feature of Additional Strokes

The strokes after the first stroke are called the additional strokes or the additional part [8]. The additional part includes a dot or a group of dots and five special symbols such as , , , , . These symbols are appeared on the top, within or under the characters and some characters have special symbols more than one. These symbols are critical to distinguish similar character contain one or more of these symbols. Fig.2 shows some examples of characters with some kinds of additional parts.



**Fig. 2.** Handwritten characters with additional parts

### 3.4 Modified Center Distance Feature-MCDF

The modified center distance feature analyzes the characters by dividing the rectangular character shape into 32 grids. The grids are unevenly divided, so that the grid units are various in sizes.

#### 3.4.1 Modified Center Distance Feature with Four Bins-MCDF-4

The MCDF-4 feature is obtained by following steps:

**Step 1:** Calculate barycentric coordinates  $P(\text{CentX}, \text{CentY})$  (gravitational center of character shape) from all character point by following expression

$$\text{CentX} = \frac{1}{n} \sum_{i=1}^{i=n} x_i \quad \text{CentY} = \frac{1}{n} \sum_{i=1}^{i=n} y_i \quad (1)$$

( n represents the total points of a character shape )

**Step 2:** The Center of gravity  $P(\text{CentX}, \text{CentY})$  divides the character's external shape into the left and right bins( $B_L$  and  $B_R$ ) by  $\text{CentX}$  at first. The left and right bins are further divided into two sub-bins respectively by  $(\text{CentX})/2$  and  $(96+\text{CentX})/2$ . Such  $96 \times 96$  rectangular shape of the character can be processed in four bins.

**Step 3:** Each part is evenly split into four units as sub-graph in vertical direction. So that 32 grids are obtained from the rectangular shape of character.

**Step 4:** Calculate barycentric coordinates  $P_j(\text{cenx}_j, \text{ceny}_j)$  (gravitational center) from the points in each grid using following expressions

$$\text{cenx}_j = \frac{1}{m} \sum_{j=1}^{j=m} x_j \quad \text{ceny}_j = \frac{1}{m} \sum_{j=1}^{j=m} y_j \quad (2)$$

(m represents the total points in a grid).

If there is not any point in the grid, the barycentric coordinate is recorded as -1

**Step 5:** Calculate the distance from the barycentric coordinate of each grid to the gravitational center of character shape by expression (3) and recorder as feature data.

$$d_j(p_j, P) = \sqrt{(cenx_j - CentX)^2 + (ceny_j - CentY)^2} \quad (3)$$

If there is not any point in the grid, the distance is recorded by the distance from the gravitational center of character shape to the origin of coordinate O(0, 0) according to the expression (4) and recorder as feature data.

$$D(O, P) = \sqrt{(CentX - 0)^2 + (CentY - 0)^2} \quad (4)$$

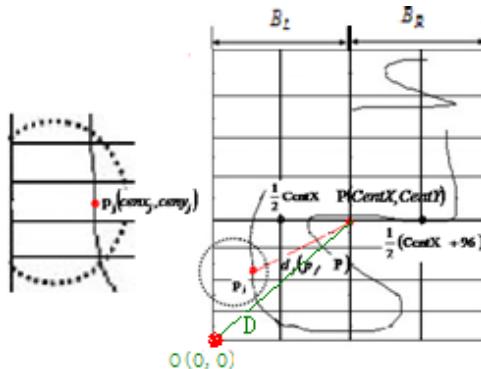


Fig. 3. MCDF-4 Feature Extraction

### 3.4.2 Modified Center Distance Feature with Two and Eight bins \_ MCDF-2 and MCDF-8

Besides the MCDF-4 feature extraction method, Modified center distance feature also can be obtained by two bins or eight bins to make grid units. The forms of Modified Center Distance features from two bins or eight bins are named MCDF-2 and MCDF-8 respectively.

### 3.4.3 Difference between MCDF and CDF

Center distance feature CDF uses the figure -1 for the grids without pen trajectory and makes those grids useless to describe characters [10], while MCDF uses all divided grids to describe the statistical and structural information of characters. MCDF lets the empty grids without pen trajectory points also describe characters by recording the location information of the character's gravitational center. Therefore, the wasted grids become useful in character recognition and make their contributions to improve the recognition rate. See Fig.3.

-1	-1	-1	-1
-1	-1	dj	dj
-1	-1	-1	dj
dj	dj	dj	dj
dj	dj	-1	-1
dj	dj	dj	-1
dj	-1	dj	-1
dj	dj	dj	-1

**Fig. 4. (a)** Feature by MCDF-4

D	D	D	D
D	D	dj	dj
D	D	D	dj
dj	dj	dj	dj
dj	dj	D	D
dj	dj	dj	D
dj	D	dj	D
dj	dj	dj	D

**Fig. 4. (b)** Feature by CDF-4

D refers the distance from the gravitational center of character shape to the origin of coordinate O(0,0), while dj refers the distance from the barycentric coordinate of each grid to the gravitational center of character shape

## 4 Experimental Results and Analysis

A total of 400 different writers contributed handwritten character samples of total Uyghur character forms ( $400 \times 128 = 51200$ ) for the experiments conducted in this paper. The system is trained using 70 percent of total samples. The remained 30 percent are participated into the identification test.

This paper uses the multi-features combination method which is performing its advantage comparing with alone using of a single feature. In Combination method, MCDF showed its advantage to recognize many similar characters and improved the recognition rate very much. In experiments, the stroke number feature and the additional part's location feature are used for pre-classification in training and testing; the shape feature and the BULR (bottom-up, left-right density ratio) features are treated as low dimensional feature (LDF) in feature combination method. CDF, MCDF are used as main feature. Recognition experiments are conducted using total 128 forms of Uyghur characters. At last, the results are compared with results from reference [11], which have been highest among previous records for total character forms with same dataset in the laboratory. The results from reference [11] are respectively from trained set which is participated in feature extraction and untrained set which only for identification test, while the results in this paper are all from untrained set. As LDF seemed quite weak in recognition without main features, only the results from features' combination are given.

**Table 1.** Average recognition rate for all Uyghur character forms (%)

Feature	CDF-4	MCDF-4	LDF and CDF-4	LDF and MCDF-4	Feature in reference[11] on trained set
Average recognition rate	73.2	67.5	77.8	87.6	75.2
Feature	CDF-8	MCDF-8	LDF and CDF-8	LDF and MCDF-8	Feature in reference[11] on un-trained set
Average recognition rate	80.7	76.7	83.2	86.4	70.7

According to the results from Table 1, we can see that:

1. In general, the feature combination method gifts higher recognition rate comparing with alone using.
2. In combination with low dimensional features, MCDF gives higher recognition rate than CDF combination.
3. MCDF-4 combination gifted a remarkable recognition rate of 87.6% and improved the recognition rate by more than 11 points compared with the result from reference [11].

MCDF gives much higher recognition rate than CDF in combination method. This is because CDF cannot use the potential of no trajectory grid units. As for the MCDF, the no trajectory located grid units make the similar characters even similar in feature domain, but put them far from the different groups. As a result, effective features for special symbols or characteristics can easily distinguish the characters within the group which consisted of similar characters in shape. However, we have to notice the similarities of character forms that even different characters may have similar writing forms, especially the middle forms of characters. The character forms with no or least special symbols are difficult to be recognized. Therefore, separate observation on the beginning form, middle form and the ending forms are highly advised.

## 5 Conclusion

This paper conducted experimental research for total forms of Uyghur basic characters. The feature combination method which MCDF is used as main feature gifted handsome recognition accuracy. MCDF proved its advantage in feature combination that recognition rate is improved by 11 points than previous record. As Uyghur character forms are different from each other, especially the middle form has some specialties which ought to carefully be considered. Separate observation and research on character forms can be the main points in future work.

**Acknowledgements.** This work is supported by the National Natural Science Foundation of China (61263038 and 61462081) and Program for new century Excellent Talents of the ministry of education (NCET-10-0969).

## References

- [1] Theodoridis, S., Koutroumbas, K.: *Pattern Recognition*, 4th edn. Academic press, USA (2009)
- [2] Shangqing, W., Fenghao, Z.: Off-line handwriting Chinese character recognition based on Bayesian grid. *Computer-Aided Engineering* 15(3), 72–74 (2006)
- [3] Zhu, B., Nakagawa, M.: *Advances in Character Recognition* (2012), under CC BY 3.0 license ISBN 978-953-51-0823-8
- [4] Tan, F.: *Online handwritten Uyghur Character Recognition Based on Mobile Platform*. Xidian University. MS thesis (2011)
- [5] Meng, Z., Zhongqiu, Y.: Image preprocessing research in Handwriting numeral recognition. *Micro-Computer Information* 22(6), 256–258 (2006)

- [6] Ranagul, D.: Research on the key technologies of online handwritten Uyghur word recognition, M.S. thesis, Xinjiang University (2011)
- [7] Jiang, X.: Feature Extraction for Image Recognition and Computer Vision. In: Proceedings of 2009 2nd IEEE International Conference on Computer Science and Information Technology, vol. 1 (2009)
- [8] Zulpiya, K.: Research on Online Uyghur Handwritten character recognition based on Feature combination. M.S. thesis, Xinjiang University (2013)
- [9] Al-Taani, A.T.: Recognition of On-line Arabic Handwritten Characters Using Structural Features. Journal of Pattern Recognition Research, 23–37 (2010)
- [10] Simayi, W., Ibrayim, M., Tursun, D., Hamdulla, A.: Research on Online Uyghur Character Recognition Technology Based on Center Distance Feature. In: Proceedings of 13th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2013), SP-6046 (2013)
- [11] Ibrayim, M., Hamdulla, A.: Design and Implementation of Prototype System for Online Handwritten Uyghur Character Recognition. Wuhan University Journal of Natural Sciences 17(2), 131–136 (2012)

# Natural Scene Text Image Compression Using JPEG2000 ROI Coding

Yuanping Zhu and Li Song

Department of Computer Science, Tianjin Normal University  
Tianjin, China  
zhuyuanping@mail.tjnu.edu.cn

**Abstract.** Regarding text region as region of interest (ROI) and assigning higher bit budget to ROIs than rest regions, ROI-based text image compression can provide both higher quality for text regions and higher compression ratio for an entire text image. JPEG2000 is a high performance image compression standard for common images but has no special optimization for text images. After image characteristic analysis, this paper proposed a natural scene text image compression method based on JPEG2000 ROI coding. ROI coding parameters are optimized for text regions. In this stage, the redundancy analysis is used to measure compression capability of different regions and scale factors in ROI coding are adjusted adaptively. With the specially designed optimization, the proposed method shows practicality in real applications. The experiment results show the improvement of compression performance which verifies the optimization effectiveness.

**Keywords:** Image compression, region of interest, natural scene text image, JPEG2000.

## 1 Introduction

Natural scene images are captured and analyzed to sense ambient environment. Text information is one kind of important content in natural scene images. Usually, natural scene images embedded by text are called natural scene text images. Natural scene text image processing, analysis, storage have gained much more focus and become a hot research area of document recognition and analysis. Especially for many applications, a large volume of natural scene text images are transmitted through internet. Decreasing image size is significant for not only saving storage but also image transmission speed-up. However, text regions have greatly different characteristics from other image content. If text regions are compressed using the same criterion with other regions, the total compression result is not so good. That maybe causes the serious degradation of text regions and the final text recognition difficulty.

Traditional document image compression technologies focus on binary document image compression, such as JBIG[1], JBIG2[2][3]. They provide great compression performance on binary document images. But natural scene text images are almost color images. For color image compression, the Joint Photographic Experts Group

(JPEG) [4] is a common still image compression standard. DCT used in JPEG is easy to cause block effect. When rising compression ratio, embedded texts are mixed with background and difficult to be discriminated. To increase the quality of compressed text images, treating text and other content respectively in compression is necessary. There are two kinds of approaches: layer-based approaches and block-based approaches. MRC (Mixed Raster Content) [5][6] is a typical layer-based approach. Which separated an image into multi-layers as text, foreground and background, then compressed different layers respectively. This is a good idea, but the real results are dependent on not only accurate text detection but also text extraction. In fact, the best of text detection accuracy in evaluation is only about 80% [8]. The situation becomes worse after combining with the text extraction. Moreover, this standard only defined decoding process. M. Thierschmann et al. proposed a raster document image compression method [7] to implement a coding mode for MRC.

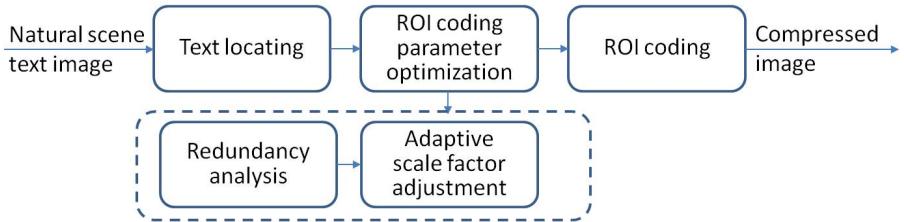
Based on wavelet, JPEG2000 [9][10] is a high performance still image compression standard. It outperforms other existing standards in general and becomes more and more popular. JPEG2000 supports ROI (Region of Interest) coding. By assigning higher bit budget to important regions than rest regions, ROI coding guarantees the higher quality of important regions while increases the compression ratio of the whole image. Since text is the most important information in natural scene text image, setting text regions as ROIs, high compression ratio with high text quality is expected. However, as a common image compression standard, JPEG2000 has no special optimization for text images. In this paper, a natural scene text image compression method based on JPEG2000 ROI coding is proposed. The ROI coding parameters are optimized for text regions. The redundancy analysis is used to measure compression capability of different regions and scale factors in ROI coding are adjusted adaptively. With the optimization, the proposed method shows practicality in real applications. Moreover, the experiment results show the improvement of compression performance which verifies the optimization effectiveness.

The rest of this paper is organized as follows. The details of the proposed natural scene text image compression method are given in Section 2. A natural scene text image compression experiment on the ICDAR2005 dataset [11] is carried out in Section 3. In the end of this paper, Section 4 draws the conclusions.

## 2 Natural Scene Text Image Compression Method

### 2.1 System Overview

The framework of natural scene text image compression system is illustrated in Fig. 1. The key stage is ROI coding parameter optimization, which carefully control the bit allocation between text regions and non-text regions. In the stage of the text locating, after text detection, ROI mask of text regions is generated for JPEG2000. To avoid too many undetected text regions, the recall priority text detection is better for this stage.



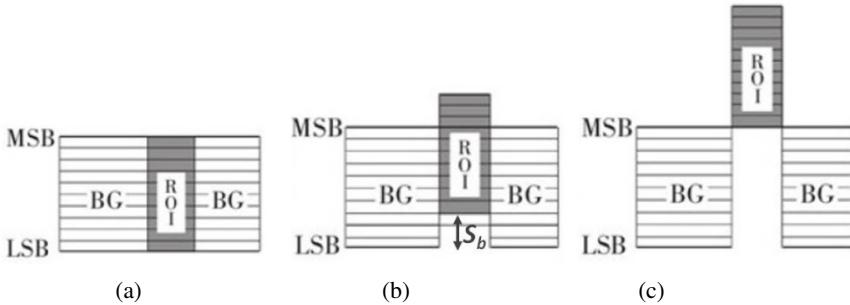
**Fig. 1.** Diagram for the natural scene text image compression system based on ROI coding

The ICDAR2005 dataset [11] is a natural scene text image set and usually used in text detection and recognition evaluation. Its training set satisfies the parameter learning of the proposed method and the experiment of this paper is carried out on this dataset. The more information of this dataset will be described in the Section 3.

## 2.2 ROI Coding of JPEG2000

ROI coding is one of the most important characteristics of JPEG2000. There are two ROI coding methods in JPEG2000: maxshift-based method given in part 1 of JPEG2000 [9] and scaling-based method given in part 2 of JPEG2000 [10]. Figure 2 illustrates two ROI coding modes. JPEG2000 is a wavelet-based image compression method. In coding process, wavelet coefficients are quantized and represented by bit planes. The bit planes are coded from the most significant bit plane (MSB) to the least significant bit plane (LSB). In lossy compression, the less significant bit planes will be discarded. That is, some small coefficients are discarded. Thus, the ROI bit planes are shifted up to achieve coding priority. The difference of two ROI coding methods is from the magnitude of shifting up. The maxshift-based method shifts up ROI coefficients beyond other coefficients to code ROIs prior to any other regions, which ensures the quality of ROIs even in a low bit rate. It supports arbitrary ROI shape without having to restore shape information. But it is easy to cause background regions not be coded at all when bit budget is insufficient. The scaling-based method utilizes bit plane shift factor to control coding priority of ROIs. In scaling based method, the ROI bit planes of quantized coefficients are shifted up by arbitrary scale factor  $S_b$  which ranges from 1 to maximum number of bit planes-12 satisfied most of applications in generally. The larger  $S_b$ , the greater the quality gap between ROIs and non-ROI regions. When  $S_b$  exceeds maximum number of bit planes, it equals to the maxshift-based method. It adjusts the bit allocation between ROIs and non-ROI regions and provides the flexibility in the quality control of ROIs. The scaling-based method only supports rectangular or elliptical ROIs in shape. But this has no negative effect for text image compression because text regions are rectangles in general.

In consideration of above characteristics, the scaling-based ROI coding is adopted in this paper. And, the scale factor is adjusted to control the quality of text regions and background. In common image compression, scale factor is usually set by empirically. For text image compression, an optimized method is designed to be adaptive to specific images.



**Fig. 2.** ROI coding methods of JPEG2000. (a) Without ROI coding; (b) Scaling-based method with scale factor  $S_b$ ; (c) Maxshift-based method.

### 2.3 Proposed Optimized ROI Coding Method

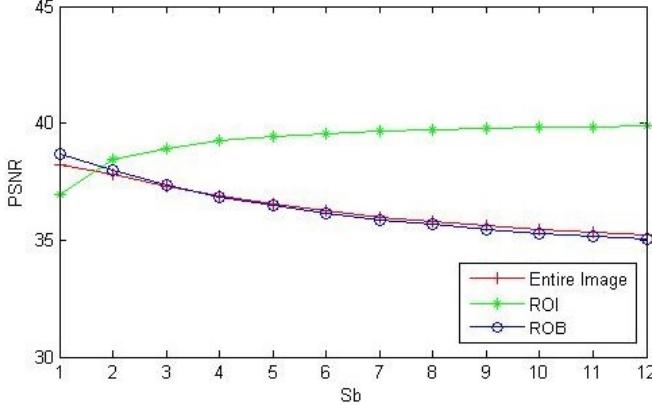
As a standard for general image compression, JPEG2000 has no optimization for specific images. To improve the compression performance for natural scene text image, an optimized ROI coding method is proposed. The first of all is to decide the optimization goal. Two key points of image compression are compression ratio and compressed image quality. For a fixed compression ratio or bit rate, the higher image quality the better. Furthermore, for ROI-based compression, the image quality should be considered in both ROIs and non-ROI regions. In natural scene text image, ROIs contain text and non-ROI regions contain general image content. In practice, text recognizer is not so as sensitive to image details as many non-text image content analyzers and human eyes. Lossy compression of text regions is acceptable and too low quality of non-text regions is not good. That means the quality of ROIs and background regions should be controlled carefully.

#### Adaptive Scale Factor Optimization

The scale factor is able to change the priority of ROI in coding sequence and adjust the quality difference between ROIs and non-ROI regions. The objective of compression optimization is to search the scale factor satisfy optimal comprehensive compression quality. Using PSNR (Peak to Signal Noise Ratio) to indicate image quality level, the optimization objective function is written as Eq. (1). Let  $P_1$  and  $P_B$  denote the PSNR values of ROI and background (non-ROI) regions.  $\gamma$  is the weight of ROI in performance evaluation. It represents the importance of the ROI image quality to the application. In general, it should be larger than 0.5. In the left part of Eq. (1), the scale factor  $S_b$  is proportional to  $P_1$  and inversely proportional to  $P_B$ . The right part of Eq. (1) has a sigmoid-like function form which is a penalty coefficient used to prevent the objective value far away from the reference range center.  $\alpha$  is set to 1.0. In Eq. (2),  $C_0$  is written as the reference PSNR rate of ROI and non-ROI regions, which is around 1.1 in this application.

$$\hat{S}_b = \arg \max_{Sb} \{ (\gamma P_I + (1-\gamma)P_B) \cdot \frac{2\alpha}{1+e^{|P_I/P_B-C_0|}} \} \quad (1)$$

$$C_0 = P_I^{ref} / P_B^{ref} \quad (2)$$



**Fig. 3.** Relation of the average PSNR and the scale factor

Figure 3 shows the relation of the average PSNR and the scale factor. All images are compressed under the reference bit rate  $b_0$ . It is observed that when the scale factor is around five to seven, the average PSNR of ROI increases smoothly and the background PSNR stays beyond 35.0db. The six is proper for the reference scale factor in this dataset. Then, the reference PSNR rate in Eq. (2) can be obtained.

There are many factors affect the compressed image quality. In fact, the real PSNR value of compressed image can only obtained after compression. To obtain the optimal scale factor, compressing the image under all scale factors and finding out the optimal one by the PSNR value is precise but impractical, because of computation cost. Here, a PSNR prediction method is utilized to estimate the PSNR value before the real compression. Besides the scale factor, several factors have close relationship with the image quality of ROI coding: *bit rate*, *ROI area*, and *redundancy*.

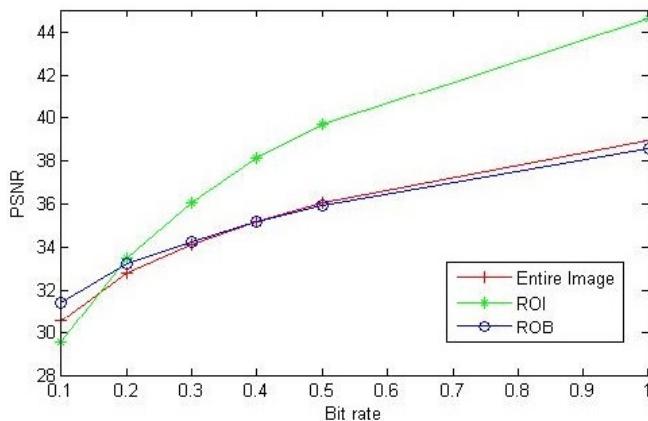
- *Bit rate*: the same to compression ratio, controlling the bit budget for the whole image, it is proportional to both  $P_I$  and  $P_B$ .
- *ROI area*: if under a same bit budget, a larger ROI means more difficult to keep the quality advantage in ROI than a small ROI. So, it is inversely proportional to both  $P_I$  and  $P_B$ .
- *Redundancy*: representing the compression potential of an image or image region, image redundancy is used to measure the possible remained quality level after compression. The redundancy is region dependent and should be calculated on ROI and background regions respectively.

Taking above factors into account, Eqs. (3) and (4) describe the relations with the PSNR values of ROIs and background regions respectively.  $b$  is the bit rate defined

by user.  $R_I$  and  $R_B$  are the redundancy of ROI and background regions. To eliminate the influence of image size, region area proportion is used to measure the ROI area factor.  $a$  and  $1-a$  are the area proportions of ROI and background region in an image respectively. When all corresponding data constructed on the training image set, multivariable regression functions  $f$  and  $g$  can be obtained. The polynomial functions are suitable here. In a practical application, a fixed reference bit rate is chosen, so the real variables of functions  $f$  and  $g$  can be reduced to 3. The average PSNR bit-rate curves are shown in Fig. 4. From that, the reference bit rate can be chosen as  $b_0=0.5\text{bpp}$ .

$$P_I = f(S_b, b, a, R_I) \quad (3)$$

$$P_B = g(S_b, b, 1-a, R_B) \quad (4)$$



**Fig. 4.** Average PSNR bit-rate curves of entire image, ROI, and background

### Redundancy Analysis

Data redundancy is a central concept in data compression. An image is equal to information plus redundancy data. The image compression is to decrease image size by reducing the redundancy while maintain information. Therefore, image redundancy is a good measurement for image compression potential. In other word, given a same quality requirement, more redundancy contained in a image means the larger probability of high compression ratio is expected and vice versa.

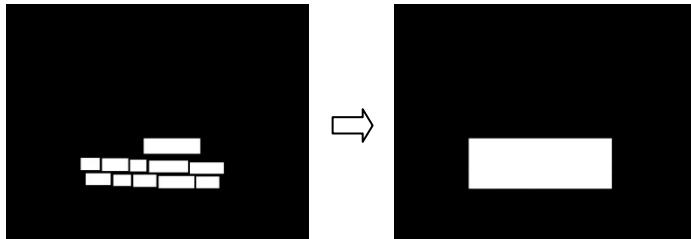
Information entropy tells how much information is in a data object. In any image, the more information, the less redundancy it contains. Information entropy of an image can be calculated as Eq. (5), where  $h_k$  is the k-th element of the image histogram. Information entropy is inversely proportional to the image redundancy. To constrain the domain, the redundancy measurement based on entropy is written as Eq. (6) which is used to compute  $R_I$  and  $R_B$  in Eqs. (3) and (4).  $\lambda$  is set to the average entropy in this paper. What should be noted here is that the histograms are constrained in corresponding regions when calculates the entropy of ROI and background regions.

$$E = -\sum_{k=1}^H p(h_k) \log p(h_k) \quad (5)$$

$$R = \frac{1}{1 + e^{-\lambda/E}} \quad (6)$$

## 2.4 Text Locating

There are two tasks in text locating: detecting text and generating text ROI mask. As mentioned before, the recall priority text detector is better. In scaling based ROI coding, multiple ROIs will degrade the coding performance [12]. Therefore, when generates ROI mask, adjacent text lines should be merged into a block to reduce ROI number. Furthermore, it is better to align the mask coordinates to code blocks of JPEG2000. Figure 5 shows the ROI mask of the sample image in Fig. 6(a).



**Fig. 5.** Text ROI mask generation

## 3 Experiments

To verify the effectiveness of the proposed method, a natural scene text image storage experiment is designed in this section. In natural scene text image analysis, text regions are important content and taken as region of interest, but background also has useful information. The ICDAR2005 dataset [11] is adopted, which has 250 training images and 249 testing images of English and Arabic number texts. A typical image sample is shown in Fig. 6. The ROI mask falls in the red rectangle. Figure 6(b) to (d) give the compressed images under bit rates 0.1, 0.3, and 0.5bpp. At 0.3bpp, text region quality becomes acceptable, but the background becomes clear until 0.5bpp.

The relation between redundancy and compression capability is the base of PSNR prediction. Figure 7 shows the data distribution between the image redundancy and the compressed image PSNR. A PSNR-Redundancy curve is fitted from 250 training images. From the observation of the curve, the redundancy has a proportional relation with PSNR. That is, the images with higher redundancy level tend to remaining higher image quality after compression, namely has greater tolerance of compression. It proves that the higher redundancy is, the larger compression capability is in an image. That is, the information entropy based redundancy measurement is effective.

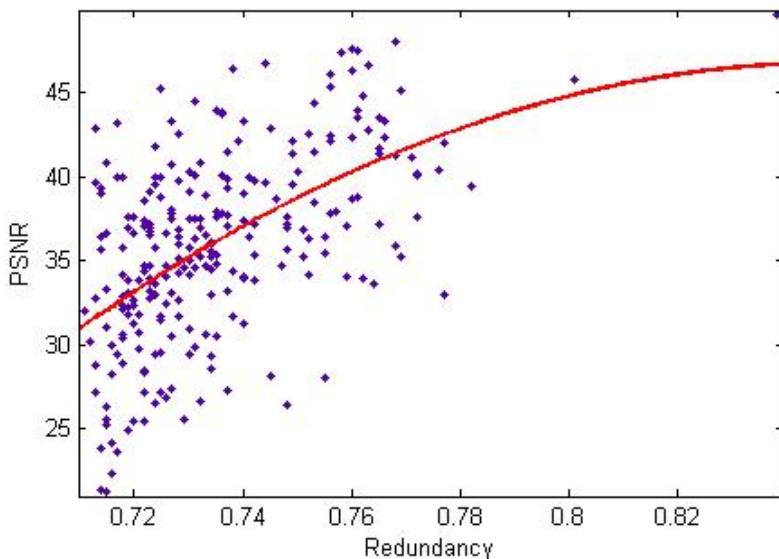
As mentioned before, to choose the reference bit rate for the application, the average PSNR bit-rate curves for ROI and background of training image set are shown in Fig. 4 respectively. Each data point in a curve is the average PSNR of all images under a same compression ratio. From this figure, it is found that text regions have higher quality than non-text regions in low compression ratio zone. When the compression ratio keeps on increasing, the situation of reverse occurs. That is a typical application dependent knowledge. For this dataset, 0.5bpp satisfies the quality requirement.

Table 1 gives the comparison of the compression on the testing set. For the better reference, the ground-truth text detection results of the dataset are used to generate text ROI masks. The second row is the result of original JPEG2000 ROI coding. The 3rd and 4th rows are the results of the proposed adaptive method.

Under the same bit rate 0.5bpp, only ROI's PSNR is a little lower than original method, but the quality of entire image is beyond it obviously. Even using a lower bit rate 0.4bpp, the proposed method's performance is still close. By approximate estimation, the proposed method with less than 0.45 bpp can obtain the similar quality with the original method under 0.5bpp. The cost is a slight drop of text region quality. Actually, the result of the proposed method is more balance between the quality of ROI and non-ROI regions.



**Fig. 6.** A typical sample in the ICDAR 2005 dataset. (a) original image; (b) to (d) compressed images with 0.1, 0.3, and 0.5 bpp.



**Fig. 7.** PSNR-Redundancy fitting curve with the bit rate 0.5 bpp

**Table 1.** Comparison of the average PSNR in the natural scene text image compression

<i>Compression Method</i>	<i>Bit rate</i>	<i>Average PSNR(db)</i>		
		<i>Text region</i>	<i>Non-text region</i>	<i>Entire image</i>
JPEG2000 without ROI coding	0.5	37.05	38.61	38.39
JPEG2000 with ROI coding	0.5	40.79	35.47	35.72
JPEG2000 with the optimized ROI coding	0.5	40.16	36.48	36.54
JPEG2000 with the optimized ROI coding	0.4	39.03	35.48	35.49

#### 4 Conclusions and Discussions

In this paper, natural scene text image compression based on JPEG2000 ROI coding is addressed. JPEG2000 is a high performance common image compression method but has no optimization for specific images. This paper proposed a ROI coding optimization method to improve the performance of JPEG2000 in text image compression. Using redundancy to measure the compression capability of image regions, the scale factors are adjusted adaptively to the images. The optimal scale factor can make full use of bit budget to obtain better comprehensive quality measurement.

The experiment verifies that the proposed ROI optimization method is effective in natural scene text image compression application. It can improve the image's comprehensive quality if under a same bit rate. Or it can provide the similar image quality with a little higher compression ratio. Actually, the proposed method can also generalized to other similar applications.

The work of this paper verified that the entropy based redundancy is effective in measuring compression potential. But it only represents coding redundancy. Considering other redundancy, such as inter-pixel redundancy will increase the completeness and the fineness of redundancy description. That will be the next research.

**Acknowledgement.** This work was supported by the National Natural Science Foundation of China (Grant No. 61203259, 61103074) and the Science and Technology Development Foundation of the Higher Education Institutions of Tianjin (Grant No.20120814).

## References

1. CCITT. Standardization of Group 3 Facsimile Apparatus for Document Transmission. CCITT Recommendation T.4 (1980)
2. ISO/IEC JTC1/SC29/WG1 N1545. JBIG2 Final Draft International Standard (December 1999)
3. Howard, P., Kossentini, F., Martins, B., Forchhammer, S., Rucklidge, W., Ono, F.: The Emerging JBIG2 Standard. *IEEE Trans. on Circuits and Systems for Video Technology* 8(5), 838–848 (1998)
4. Information Technology - Digital Compression and Coding of Continuous-Tone Still Images - Requirements and Guidelines, ITU-T Recommendation T.81 - ISO/IEC 10918-1 (1992)
5. Mixed Raster Content (MRC), ITU-T Recommendation T.44 (2005)
6. Cheng, H., Bouman, C.A.: Document Compression Using Rate-Distortion Optimized Segmentation. *Journal of Electronic Imaging* 10(2), 460–474 (2001)
7. Thierschmann, M., Bartel, K., McPartlin, S., Martin, U.: New Technology for Raster Document Image Compression. In: Proc. of SPIE 3967, Document Recognition and Retrieval VII, vol. 286 (December 1999)
8. Karatzas, D., Shafait, F., et al.: ICDAR 2013 Robust Reading Competition. In: Proc. of the 12th International Conference on Document Analysis and Recognition, pp. 1115–1124 (2013)
9. ISO/IEC 15444-1. JPEG2000 image coding system -Part 1: core coding system. Tech. Report, ISO (2000)
10. ISO/IEC JTC1/SC20 WG1 N2000. JPEG2000 Part 2: final committee draft. Tech. Report, ISO (2000)
11. Lucas, S.M.: ICDAR 2005 text locating competition results. In: Proc. of the 8th International Conference on Document Analysis and Recognition, pp. 80–84 (2005)
12. Subedar, M.M., Karam, L.J., Abousleman, G.P.: JPEG2000-Based Shape Adaptive algorithm for the Efficient Coding of Multiple Regions-of-Interest. In: Proc. of 2004 International Conference on Image Processing, vol. 2, pp. 1293–1296 (2004)

# Off-Line Uyghur Handwritten Signature Verification Based on Combined Features

Kurban Ubul<sup>1</sup>, Tuergen Yibulayin<sup>1,\*</sup>, and Alimjan Aysa<sup>2</sup>

<sup>1</sup>School of Information Science and Engineering, Xinjiang University  
830046 Urumqi, China  
[{kurbanu,Tuergen}@xju.edu.cn](mailto:{kurbanu,Tuergen}@xju.edu.cn)

<sup>2</sup>Center of Network and Information Technology, Xinjiang University  
830046 Urumqi, China  
[{alim}@xju.edu.cn](mailto:{alim}@xju.edu.cn)

**Abstract.** An off-line Uyghur handwritten signature verification method based on combined features was proposed in this paper. Firstly, the signature images were preprocessed using techniques adapted to the Uyghur signature. The preprocessing included noise reduction, binarization, and normalization. Then, the global features, local features which each of them include several features were extracted respectively after the preprocessing, and they are combined together. Finally, two types of classifiers, Euclidean distance classifier, and non-linear SVM classifier are used to classify 75 genuine signatures and 36 random forgeries in our experiment. Two kinds of experiments were performed for and variations in the number of training and testing datasets. Experiments indicate that the combination of directional features with local central point features has obtained 2.26% of FRR and 2.97% of FAR with SVM classifier. The experimental results indicated that the combination method can capture the nature of Uyghur signature and its writing style effectively.

**Keywords:** Uyghur, handwritten signature, combined features, verification.

## 1 Introduction

Handwritten signatures are the most widely accepted biometric to human identity recognition and verification all around the world both socially and legally [1]. The aim of the signature verification process is to confirm or reject the sample that it is belong to 1:1 classification problem [2]. Signature verification can be divided in two classes: on-line and off-line signature verification. On-line (or dynamic) systems use a digitizer or an instrumented pen to generate signals; while off-line (or static) systems produce an image of a signature with the help of a camera or scanner [3, 4].

Surveys of the state of the art off-line signature recognition and verification systems designed up to 1993 appear [3]. Another survey paper [1] had summarized the approaches used for off-line signature recognition and verification from 1993 to 2000. An off-line Arabic signature recognition and verification system was proposed in [2].

---

\* Corresponding author.

Its recognition phase was used the multi-stage classifier and a combination of global and local based features whereas the verification was done using fuzzy concepts. Dakshina et al [5] used SVM to fuse multiple classifiers for an off-line signature system. Lv et al used from both static features and dynamic features and SVM as classifier to verify the Chinese signatures [6]. Neural network based signature recognition and verification methods indicated in [5, 7]. Euclidean distance classifier based signature verification proposed in [8]. Xiao et al [9] proposed a way of off-line Chinese handwritten recognition based on wavelet packs and Gauss model in order to solve the problems of efficient feature extraction. Dynamic Time Warping (DTW) based signature verification system proposed in [10]. It works by extracting the vertical projection features from the signature images, and then comparing it to a reference. An approach for off-line Persian signature identification and verification was proposed in report [11] that was based on image registration, discrete wavelet transform and image fusion. The grayscale image features based offline handwritten signature verification technique using the histogram displacement proposed in reference [12], and the co-occurrence matrix and local binary pattern is used to extract the features. These reports about signature recognitions were mostly based on Latin handwriting [1, 3, 5, 7], Chinese handwriting [5, 6], Arabic handwriting [2, 7] and Persian handwriting [11]. However, there are only two reports that were our previous research for off-line Uyghur signature recognition. Modified grid information features based Uyghur signature recognition proposed in [13], and thinning effects on the accuracy of Uyghur signature recognition are studied in [14], and no reports about Uyghur handwritten based signature verification. So there is a great need and much research space for implementing existing algorithms creatively or developing new effective algorithms suitable to the nature of Uyghur handwritten signature.

In this paper, multi-features including global features and local features extracted based on the nature of Uyghur signature, and they are combined together. Experiments were performed using Euclidean distance classifier, and non-linear SVM classifier for Uyghur signature samples of 75 genuine signatures, 36 random forgeries.

## 2 Data Acquisition and Preprocessing

Common steps of signature recognition include data acquisition, pre-processing, feature extraction and classification. Uyghur people are selected to give their natural signatures on paper for data acquisition. The signature image must be preprocessed to reduce noise and account for different sizes.

### 2.1 Data Acquisition

The signatures were collected using black ink from 380 Uyghur people, on a white A4 sheet of a paper, with 21 signatures per page. Each paper was divided into 21 same sized boxes into which the person was asked to sign his/her signature. The area of each box is big enough to give enough space to signer, and to allow size deviation of signature. The signatures are digitized using Canon MP810 scanner, and stored with

300 dpi resolution and .bmp format in 256 grey levels. For the signature verification issue, 50 forgery samples (21 signatures /sample) are written.

## 2.2 Pre-processing

For the Uyghur signature images, the pre-processing steps include noise reduction, binarization, and size normalization and thinning. Since pre-processing is not focus in this paper and it is adapted to the Uyghur signature in our previous work [13, 14]. Since the pre-processing steps for Uyghur signature verification in this paper are same as the [13, 14], so it is not explained here.

## 3 Feature Extraction

Two types of features such as global and local features were extracted in this paper.

### 3.1 Local Features

The local features extracted in this paper include directional features (DF) and local central point (LCP) features.

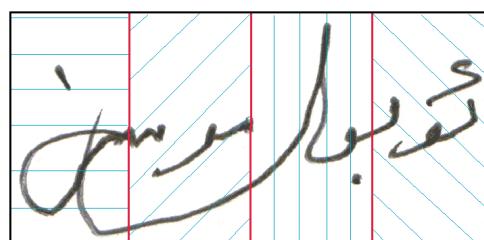
1) The directional feature (DF): The directional feature proposed in this paper is a kind of statistical feature. For the extraction of this feature, the signature image is vertically divided in to 4 same sized parts at first. Each part of the signature is scanned in  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  direction separately as indicated in Figure1. Then, 4 dimensional features are extracted after counting the numbers of black pixels in each direction separately. It is taken 16-dimention directional features finally. The directional feature vectors extracted from a signature image is:

$$F_{p^j} = [f_j^1, f_j^2, \dots, f_j^d] \quad (1)$$

where,  $j$  is the training samples for a person's signature, in this paper  $j=1, 2, \dots, 19$ ,  $d$  is the dimension of features with value of 16,  $p$  indicated class type of sample (person). The directional feature vectors for all the training samples are:

$$F = [F_1^j, F_2^j, \dots, F_n^j] \quad (2)$$

where,  $n$  is the numbers of training samples.



**Fig. 1.** Directional features

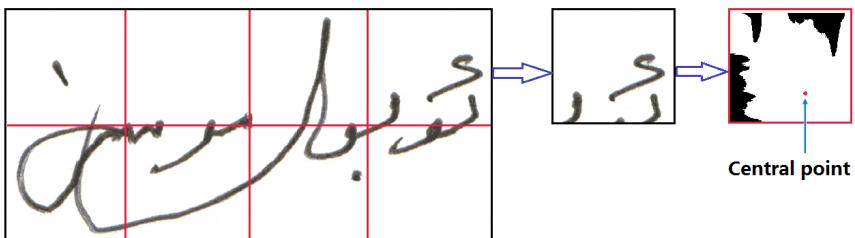
2) The local central point features: It was extracted 32 dimensional local central point features to combining with the nature of Uyghur signature. The signature segmentation pattern of local central features indicated in [2] was modified, and each signature image was divided into  $2 \times 8$  rectangular area. Then, the central point of each grid was calculated via horizontal and vertical projection profile separately. The abscissa and ordinate of each central point is taken as feature, the 32 -dimensional features were extracted in this way.

$$\begin{cases} P_h[y] = \sum_{x=1}^m \text{black.pixel}[Z(x, y)] \\ P_v[x] = \sum_{y=1}^k \text{black.pixel}[Z(x, y)] \end{cases} \quad (3)$$

where, *black pixel* is refers to the amount of black pixels for signature, where,  $x=1, 2, \dots, k$ ;  $y=1, 2, \dots, m$ . The central point of each window:

$$\begin{cases} C_h = \sum_{y=1}^m (y \cdot p_h[y]) / \sum_{y=1}^m p_h[y] \\ C_v = \sum_{X=1}^K (x \cdot p_v[x]) / \sum_{x=1}^k p_v[x] \end{cases} \quad (4)$$

where,  $C_h$  and  $C_v$  is the abscissa and ordinate of central point. A central point of a grid in Uyghur signature indicated as the following Figure 2.



**Fig. 2.** The local central point features

### 3.2 Global Features

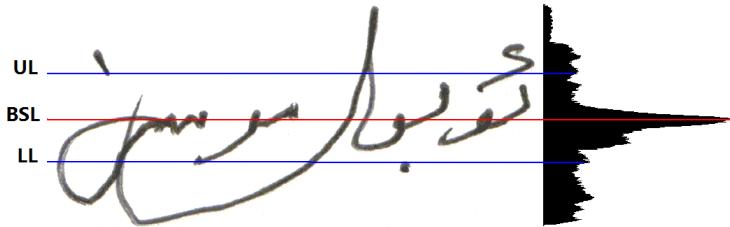
The global features extracted here include global base line (BSL), upper line (UL) and lower line (LL) features, and modified grid information (MGI) features.

(1) The global baseline features: The global baseline indicated to the maximum point of the smoothed global vertical projection curve [2] as indicated in the following Figure 3. If the size of each signature image is  $m \times k$ , the global baseline  $P_m$  is:

$$P_m = \max\{p_v[x]\} \quad (5)$$

where,  $P_v[x]$  is vertical projection of the image that indicated in equation (3).

(2) The upper and lower line features: The upper and lower line which are 20 pixel far (up or down) from the BSL were also extracted to using same projection above and under the baseline separately indicated as the following Figure 3.



**Fig. 3.** Base line, upper & lower line

(3) Modified grid information features: Grid information features are one of the common features in off-line signature recognition and verification [5]. It is divided into many rectangular parts (usually  $12 \times 8$ ,  $15 \times 8$ ,  $10 \times 8$  parts), and the sum of foreground pixels is calculated for each part. In general, the rectangle with the smallest number of black pixels is taken as zero, and the rectangle with the highest number of black pixels is taken as one. For Uyghur signature, it is respectively segmented into  $8 \times 8$  rectangular parts in horizontal and vertical direction. Then, the signature image was scanned in 4 direction, such as left to right, right to left, top to down and down to top. Since, the feature extraction way is similar in the 4 directions, and MGI features are explained in detail in our previous work in [13], so it is explained to take feature extraction method in left to right direction as example.

The signature image is segmented horizontally to 8 parts. The first feature in each grid is to be set as the sum of white pixels calculated from right point of black pixel firstly met with initially in the right-left direction. Next, the second feature is extracted as that the sum of the white pixels calculated between first and second block of black pixels in same direction. In this way, 16 features are extracted in the left to right direction from these grids. It is also extracted 16 dimensional features in other 3 directions in the same way, thus, 64-dimentional MGI features were extracted here.

## 4 Classification

The Euclidean distance classifier [8] and non-linear Support Vector Machine (SVM) classifier are used in verification stage. The non-liner SVM classifier is to be selected here since it has more accuracy than liner one and other. If the weight vector can be expressed as a linear combination of the training examples, i.e.  $w = \sum_{i=1}^n a_i x_i$ , then:

$$f(x) = \sum_{i=1}^n a_i x_i^T x + b \quad (6)$$

In the feature space F, this expression takes the form:

$$f(x) = \sum_{i=1}^n a_i \phi(x_i)^T \phi(x) + b \quad (7)$$

The representation in terms of the variables  $\alpha_i$  is known as the dual representation of the decision boundary. The feature space F may be high dimensional as indicated equation (7), making this trick impractical unless the kernel function  $k(x, x')$  can be computed efficiently denoted as:

$$k(x, x') = \phi(x)^T \phi(x') \quad (8)$$

In terms of the kernel function the discriminant function is:

$$f(x) = \sum_{i=1}^n a_i k(x, x_i) + b \quad (9)$$

## 5 Experimental Results

Two types of classifiers, Euclidean distance classifier, and non-linear SVM classifier are used in our experiment. 75 genuine signatures, 36 random forgeries which are contained by genuine signatures are selected from our Uyghur handwritten database. Two kinds of experiments have been carried out according to the numbers of training samples. We trained 36 and 48 genuine signatures during the experiments respectively. The testing and training procedure was repeated 6 times with different randomly chosen training and testing sets for the get more reliable results in each case.

Usually, efficiency of signature verification results are reported in terms of False Acceptance Rate (FAR), which means a forgery signature is considered as a genuine one, False Rejection Rate (FRR), which means a genuine signature is considered as a forgery. The verification system can obtain lower percentage both FRR and FAR is acceptable and indicate its strong efficiency.

The experimental results using Euclidean distance classifier and non-linear SVM classifier are indicated in table 1, table 2 as below.

It is clear from the Table 1 that Euclidean distance classifier indicates lower verification results with LCP features that it's FRR and FAR is about 9.86% and 12.35% respectively. They are decreased with the combination of DF and LCP features that 3.97% of FRR and 3.22% of FAR. The experimental results indicate that the method which combining DF and LCP feature is more efficient than others.

SVM classifier indicates higher verification results than Euclidean distance classifier with LCP features that it's FRR and FAR is about 7.56% and 9.82% respectively. It is declined to 5.61% and 8.24% when the numbers of training datasets are increased to 48. They are further decreased with the combination of DF and LCP features that 2.26% of FRR and 2.97% of FAR. Others are illustrated in the Table 2.

**Table 1.** Signature verification results (percentage) with the euclidean distance classifier

Features	Numbers of training genuine signature			
	Training 36 samples		Training 48 samples	
	FRR (%)	FAR (%)	FRR (%)	FAR (%)
LCP features	9.86	12.35	8.23	10.79
Directional features	8.51	6.74	6.44	5.16
MGI features	6.07	9.29	3.96	7.68
BSL & LCP	7.43	10.1	5.35	8.72
BSL & DF	6.87	4.93	5.28	3.64
BL & UL & LL & DF	5.92	4.56	4.72	3.53
DF & LCP	5.38	3.81	3.97	3.22

**Table 2.** Signature verification results (percentage) with the non-linear svm classifier

Features	Numbers of training genuine signature			
	Training 36 samples		Training 48 samples	
	FRR (%)	FAR (%)	FRR (%)	FAR (%)
LCP features	7.56	9.82	5.61	8.24
Directional features	7.78	5.91	6.03	4.35
MGI features	5.63	7.46	3.39	5.82
BSL & LCP	6.74	8.53	4.45	6.90
BSL & DF	6.27	4.38	4.79	3.31
BSL & UL & LL & DF	5.15	4.11	4.23	3.26
DF & LCP	4.38	3.58	2.26	2.97

It can be seen from the two tables (Table 1, Table 2) that the method which combining directional and local central point features is more efficient than other feature extraction methods, and it can capture the writing style of Uyghur signature more efficiently.

## 6 Conclusions and Future Work

In this paper, an off-line signature verification system for Uyghur handwriting was presented. Signature images were preprocessed according to the nature of Uyghur signature. Then global and local features which each of them include several features were extracted based on the structure of Uyghur signature, and they are combined together. Euclidean distance classifier and non-linear SVM classifier are used in our experiment. Two kinds of experiments were performed for and variations in the number of training and testing datasets. Experiments indicate that the combination of directional features with local central point features has obtained 2.26% of FRR and 2.97% of FAR with SVM classifier. And its verification results are higher than others in this paper. Experimental results indicated that this kind of combination method can capture the nature of Uyghur signature and its writing style.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (No. 61163028, 61363064), Special Training Plan Project of Xinjiang Uyghur Autonomous Region's Minority Science and Technological Talents (No. 201323121), College Scientific Research Plan Project of Xinjiang Uyghur Autonomous Region (No. XJEDU2013I11) and the Open Projects Program of National Laboratory of Pattern Recognition (No. 201306321).

## References

1. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: a comprehensive survey. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22(1), 63–84 (2000)
2. Ismail, M.A., Gad, S.: Off-line Arabic signature recognition and verification. *Pattern Recognition* 33(10), 1727–1740 (2000)
3. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification – the state of the art. *Pattern Recognition* 22(2), 107–131 (1989)
4. Guru, D., Prakash, H.: Online Signature Verification and Recognition: An Approach Based on Symbolic Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(6), 1059–1073 (2009)
5. Huang, K., Yan, H.: Off-line Signature Verification Based on Geometric Feature Extraction and Neural Network Classification. *Pattern Recognition* 30(1), 9–17 (1997)
6. Lv, H., Wang, W., Wang, C., Zhuo, Q.: Off-line Chinese signature verification based on support vector machines. *Pattern Recognition Letters* 26(15), 2390–2399 (2005)
7. Kisku, D.R., Gupta, P., Sing, J.K.: Off-line signature identification by fusion of multiple classifiers using statistical learning theory. *International Journal of Security and Its Applications* 4(3), 35–45 (2010)
8. Yingyong, Q., Hunt, B.: Signature Verification Using Global and Grid Features. *Pattern Recognition* 22(12), 1621–1629 (1994)
9. Baltzakisa, H., Papamarkos, N.: A new signature verification technique based on a two-stage neural network classifier. *Engineering Applications of Artificial Intelligence* 14(1), 95–103 (2001)
10. Shanker, A., Rajagopalan, A.: Off-line signature verification using DTW. *Pattern Recognition Letters* 28(12), 1407–1414 (2007)
11. Ghandali, S., Moghaddam, M.E.: Off-line Persian signature identification and verification based on image registration and fusion. *Journal of Multimedia* 4(3), 137–144 (2009)
12. Vargas, J., Ferrer, M., Travieso, C., Alonso, J.: Off-line signature verification based on grey level information using texture features. *Pattern Recognition* 44(2), 375–385 (2011)
13. Ubul, K., Adler, A., Abliz, G., Yasin, M., Hamdulla, A.: Off-Line Uyghur Signature Recognition Based on Modified Grid Information Features. In: The 11th International Conference on Information Sciences, Montreal, Canada, July 3–5 (2012)
14. Ubul, K., Adler, A., Yadikar, N.: Effects on Accuracy of Uyghur Handwritten Signature Recognition. In: Liu, C.-L., Zhang, C., Wang, L. (eds.) CCPR 2012. CCIS, vol. 321, pp. 548–555. Springer, Heidelberg (2012)

# Off-Line Signature Verification Based on Local Structural Pattern Distribution Features

Wen Jing, MoHan Chen, and JiaXin Ren

College of Computer Science, Chongqing University, Chongqing, 400044

**Abstract.** Handwritten signature is a widely used biometric. The most challenging problem in automatic signature verification is to detect skilled forgery which is similar to the genuine signatures. This paper presents a novel method for extracting features for off-line signature verification. These features is based on probability distribution function, which characterizes the frequent structural patterns distribution of a signature image. Experiments were conducted on an publicly available signature database MCYT corpus. Experimental results show that the proposed method was able to improve the verification accuracy.

**Keywords:** Off-line signature verification, Pattern recognition, Local structural pattern, Chi-square distance.

## 1 Introduction

With the increasing security requirements of todays society, biometrics is playing a more and more important role. As one of the oldest biometrics, signature is the result of rapid human movements depending on the psychophysical state of the signer and the signing conditions. The signature verification system performs one-to-one and determines whether the two samples of handwriting were written by the same person.[17] Approaches to signature verification fall into two categories: on-line and off-line [14]. Even today, high success rates are still limited to the on-line.[1] This is because on-line signature verification can capture dynamic features like time, pressure, speed and the order of stroke. However, off-line verification is more user friendly and have a significant advantage in many of the practical uses since they do not require access to special device. Up to now, off-line signature verification still an open research area needed more efforts to address it. In signature verification system, three kinds of forgery may be considered: random forgery, simple forgery, and skilled forgery[4]. Naturally the skilled forgery is very similar to the genuine signatures and is more difficult to be distinguished, especially for off-line signature verification due to the lack of dynamic information, so skilled forgery detection is the most challenging job for off-line signature verification[10].

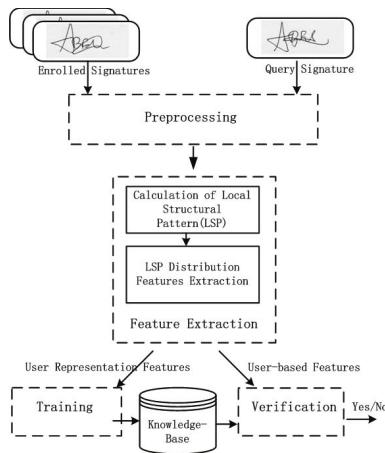
During the last few years, researchers have tried different methods with various approaches to detect the skilled forgeries detection. An extensive overview of previous work is included in [14,3]. J. F. Vargas et al. [7] proposed an off-line signature verification system based on grey level information using texture

features. They adopted the co-occurrence matrix and local binary pattern as features. In [13], surroundedness feature is proposed, which contains both shape and texture property of a signature. K.Tselios et al.[9] proposed grid-based feature distributions, this method explored the relative pixel distribution along a signature trace.

In this paper, we mainly present new and very effective techniques for signature verification that use probability distribution functions extracted from the scanned images of handwriting to characterize signer individuality. This paper is organized as follows: Section 2 presents the preprocessing and the feature extraction methodology. Section 3 shows the experimental results based on MCYT corpus. Finally ,conclusions are discussed in section 4 .

## 2 Methodology

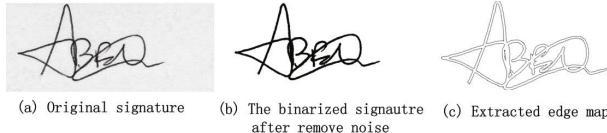
In order to perform verification of a signature, several steps must be performed. Figure 1 illustrates the whole signature verification process. Initially the scanned signature image is preprocessed. The out image is used to extract features. Finally signature is verified by matching extracted features against those stored in the database.



**Fig. 1.** Block diagram of the proposed verification system

### 2.1 Preprocessing

Some preprocessing steps have to be applied to the input signature images. the signature images are first binarized using the OSTU algorithm [11]. And then, we apply mathematical morphology method is used to remove the noise of small area. Finally, edge detection based on Sobel operator was performed on each signature image. Figure 2 shows an example of processed signature image.



**Fig. 2.** Sample of original and after-preprocessing signature

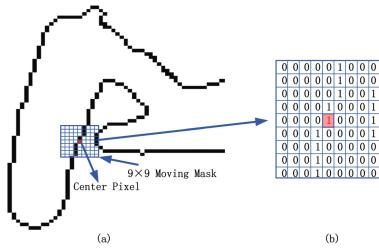
## 2.2 Feature Extraction

Similar to many other pattern recognition problems, feature extraction is a crucial step. In off-line signature verification , an efficient feature extraction technique should adequately describe the information of signature and could be tolerant to intra-user variability. Additionally, in order to detect skilled signature where forgeries are visually much similar to the genuine signatures on a global scale, local measurement to extract pertinent detailed information are needed. In this paper, we use local structural features to uniquely characterize a candidate signature, and these features characterizes the frequent structural pattern distribution, and the steps are as follows:

The first step of generating the new feature is extracting segments block of signature. To obtain these patterns, we use a  $n \times n$  sliding window that is slid over an edge-detected binary handwriting image, and for each sliding window, the central pixel is on the edge pixel. The size  $n$  of window is even, and should be large enough to contain ample information about the style of the writer and small enough to ensure a good identification performance[15].Regarding the mask size  $n$ , we carry out an exhaustive study with sizes of  $7 \times 7, 9 \times 9, 11 \times 11, 13 \times 13$  for our system. The effect of mask size on verification performance has been analyzed in detail in Section 3. The method has been illustrated in figure 3.

For simplify, we employ a part of signature as sample in figure 3. In the sliding window, the numbers 1 and 0 represent edge pixels and non-edge pixels respectively, and the red pixel represents the center edge pixel. each window is a segments block of signature, and the number of the segments block is equal to the number of the edge pixels for a signature image.

After obtaining the segments block, we need model each segments block through encoding. each segments block mainly provides of two-part information including shape information of the main segment and structural information of different segments, as shown in figure 4, different numbers represent the different connected domains. Every segment is a connected domain in sliding window, and the main segment is including the center pixel. we could describe the whole signature through coding the main segments. In addition, there is more than one connected domain in each sliding window, and structural information between different connected domains also need to be described. So, we model local patterns from two aspects including the main segment and the different segments. For the main segment in the sliding window, all pixels except for the central pixel could be divided into multiple groups according to the different Chebyshev-distance [2] between the central pixel and other pixels. All pixels



**Fig. 3.** Extracting segments block of signature(a)Sample of edge-detected binary signature image with a sliding window,(b)A segments block

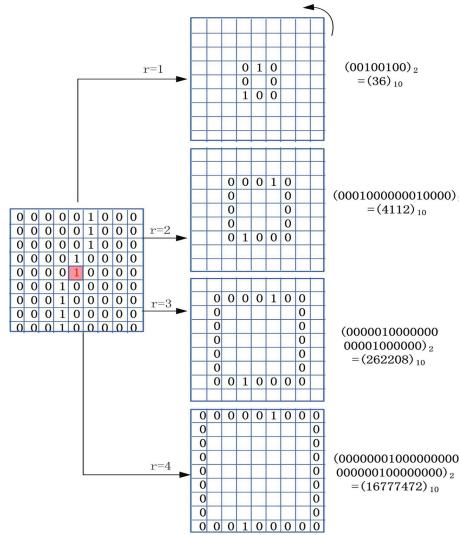
0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 1 0 0 2
0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 1 0 0 2
0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 1 0 0 2
0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 1 0 0 2
0 0 0 0 0 1 0 0 0 1	0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 1 0 0 2
0 0 0 0 1 0 0 0 0 1	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 2
0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0	0 0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0 0

(a) (b) (c)

**Fig. 4.** (a)A local structural pattern,(b)The main segment,(c)multiple segments

with distance  $r$  form a binary sequence in accordance with a certain order such as counter-clockwise, and each binary sequence finally produces a decimal value  $lm$ . That is to say, each  $lm$  is a pattern which denotes the shape information of the main segment and we use LSPM to stand for these patterns. Additionally, it is worth noting that we only focus on pixel set containing at least two 1 to reduce the impact of noise. For example, it can be seen from 5 that around the center pixel, the size of sliding window is 7, the decimal value at distance 1 is 36, at distance 2 is 4112, at distance 3 is 262208, and at distance 4 is 16777472.

For multiple segments, we use the similar way to code. But, we only consider pixel sequences including different connect domains in order to represent the structural relation between different segments. As shown in figure6, for distance  $r = 1, 2, 3$ , these pixels from the same connected domain, so these pixel sequences are ignored and only the pixel sequence with distance 4 are encoded. For simplicity, the experiments below use a coding scheme that splits each pixel sequence into multiple binary sequence, and each binary sequence contain only two number 1 which are from two different connected domains. Thus, each pixel sequence maybe produce multiple decimal values  $ls$ . As illustrated in figure6. Here, we use LSPS to represent the structural information between the different segments.



**Fig. 5.** Code for the main segment

Through above steps , all local structural patterns(LSPs) which consist of LSPM and LSPS are obtained. The third step is creating LSP distributions (LSPD). Therefore, in this step, we calculate frequency of each LSPM and LSPS respectively, which is given by:

$$H1_r(lm) = \frac{LSPM_r(lm)}{\sum_{r=1}^{(n-1)/2} LSPM_r(lm)} \quad (1)$$

$$H2_r(ls) = \frac{LSPS_r(ls)}{\sum_{r=1}^{(n-1)/2} LSPS_r(ls)} \quad (2)$$

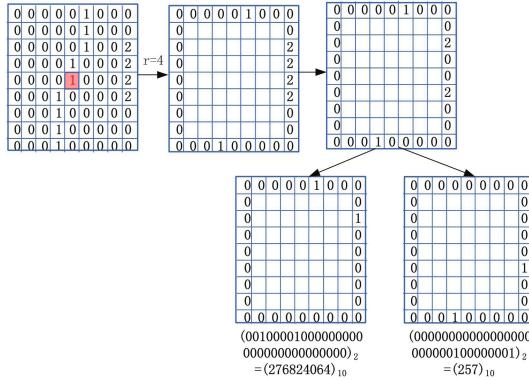
where  $r$  corresponds to distance between center pixel and around pixels in sliding window, and  $n$  is the size of sliding window.i.e., if  $7 \times 7$  size window is used then  $n$  is 7. $LSPM_r(lm)$  is the number of LSPM pattern  $lm$  at distance  $r$ .Similarly, $LSPS_r(ls)$  is the number of LSPS pattern  $ls$  at distance  $r$ .

Therefore, the LSPD is defined as

$$LSPD = \{H1_r(lm), H2_r(ls)\} \quad (3)$$

## 2.3 Classification

The above described features are extracted from a sample group of signature images of different person. In the classification phase, various classifiers have

**Fig. 6.** Code for the different segments

been exploited to authenticate handwritten signatures. In this work, we use Chi square distance, which is one of well-known goodness-of-fit statistics[16], is adopted:

$$\chi^2 = \sum_{s=1}^S \sum_{l=1}^{L_s} \frac{(S_{sl} + M_{sl})^2}{(S_{sl} - M_{sl})} \quad (4)$$

where  $S$  is the number of scales and  $L_s$  is the number of LSP pattern types on scale  $s$  and  $S_{sl}$  and  $M_{sl}$  correspond to the sample and model probabilities at pattern  $l$  on scale  $s$ , respectively.

During verification, a claimed signature is compared against our template file using Chi square distance and if it is below a certain threshold value, then this signature is accepted as genuine, otherwise it is rejected to be a forgery. Here, we use localized thresholds. What this means is that each signature that is stored in the template will be stored with its own unique threshold.

### 3 Experiment and Results

Experiments are conducted on the publicly available signature database MCYT corpus. We adopt AER(Average Error Rate), FAR(False Acceptance Rate) and FRR(False rejection rate) to evaluate the verification performance.

#### 3.1 Signature Database

MCYT is a bimodal database used for the experiments. [12] Off-line signature subcorpus comprises 2250 signature images, with 15 genuine signatures and 15 forgeries per user (contributed by three different user-specific forgers). The 15 genuine signatures were acquired at different times (between three and five) of

the same acquisition session. At each time, between 1 and 5 signatures were acquired consecutively.[6]

In the training and testing phase, the genuine (for threshold calculation) training samples will be chosen randomly from the database set and the test will be performed with the other genuine and forged samples. In order to obtain reliable results in each studied case, the training and test procedure was repeated 10 times with different randomly chosen training sets. In our experiment, 10 genuine samples were used for training, five genuine samples and 15 forgery samples were used for testing.

### 3.2 Experimental Results on Different Window Sizes

The selection of the sliding window size would directly affect verification performance.

**Table 1.** Experiment results with different window sizes on MCYT corpus

Window size	FAR(%)	FRR(%)	AER(%)
$7 \times 7$	8.72	14.94	11.83
$9 \times 9$	6.58	14.4	10.49
$11 \times 11$	3.56	14.94	9.25
$13 \times 13$	4.8	15.2	10

From Table 1 , we can find that error rate will change with the number of window size increases. This is because that if the sliding window size is too small, the description of local structural pattern is not comprehensive, because no two signatures by the same person are identical on a detailed scale, and error rate will be increased. On the contrary, if the sliding window size is too large, the set of features would contain much redundant information, and the error rate would also be increased.

### 3.3 Comparison with Some Other Published Methods

The lack of a standard international signature database, so a comparison of the performance of different signature verification systems is a difficult task. For the sake of completeness, we present some results obtained by published studies that used the MCYT corpus in Table 2.

**Table 2.** Performance comparison of the proposed methods with other published methods

Method	AER(%)
[5]	22.4
[8]	15.02
[7]	11.28
proposed method	9.25

## 4 Conclusion

In this work, a feature extraction technique for analysing the handwritten signature for verification tasks is proposed. The method is based on the idea of statistically exploiting the relative local structural pattern distribution by the sliding window. Verification is based on Chi-square distance classification algorithm. Experimental results demonstrated that the proposed method achieved favorable verification performances. Further work is expected to be carried out towards the study of the selection of features and other handwritten datasets.

**Acknowledgments.** This work is supported by the Program for Natural Science Foundations of China(61103116, 61173129, 61100114) and the Fundamental Research Funds for the Central Universities(106112013CDJZR180002).

## References

1. Kovari, B., Charaf, H.: A study on the consistency and significance of local features in off-line signature verification. *Pattern Recognition Letters* 34(3), 247–256 (2013)
2. Cantrell, C.D.: Modern mathematical methods for physicists and engineers. Cambridge University Press (2000)
3. Impedovo, D., Pirlo, G.: Automatic signature verification: the state of the art. *IEEE Trans.Syst.Man Cybernet. Part C* 38(5), 609–635 (2008)
4. Justino, E.J.R., Bortolozzi, F., Sabourin, R.: An off-line signature verification using hmm for random,simple and skilled forgeries. In: Sixth Intl. Conference on Document Analysis and Recognition(ICDAR), pp. 1031–1034 (2001)
5. Fernandez, F.A., Fairhurst, M.C., Fierrez, J., Ortega-Garcia, J.: Automatic measures for predicting performance in off-line signature verification. In: Proc.Int. Conf. on Image Processing, pp. 369–372 (2007)
6. Fierrez-Aguilar, J., Alonso-Hermira, N., Moreno-Marquez, G., Ortega-Garcia, J.: An off-line signature verification system based on fusion of local and global information. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 295–306. Springer, Heidelberg (2004)
7. Vargas, J.F., Ferrer, M.A., Travieso, C.M., Alonso, J.B.: Off-line signature verification based on grey level information using texture features. *Pattern Recognition* 44, 375–385 (2011)
8. Wen, J., Fang, B., Tang, Y., Zhang, T.P.: Model-based signature verification with rotation invariant features. *Pattern Recogntiion* 42, 1458–1466 (2009)
9. Tselios, K., Zois, E.N., Nassiopoulos, A., Economou, G.: Grid-based featurer distributions for off-line signature verification. *IET Biometrics* 1(1), 72–81 (2012)
10. Batista, L., Rivard, D., Sabourin, R., Granger, E., Maupin, P.: Pattern recognition technologies and applications:recent advances. In: IGI Global snippet,CH.III:State of the Art in off-line Signature Verification pp. 39–62 (2008)
11. Otsu, N.: Athreshold selection method from gray-level histogram. *IEEE Trans. Syst. Man Cybernet.* 6, 62–66 (1979)
12. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez- Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.J., Vivaracho, C., Escudero, D., Moro, Q.I.: Mcyt baseline corpus: a bimodal biometric database. *IEE Proc.Vis. Imag. Sign. Process.* 150(6), 395–401 (2003)

13. Kumar, R., Sharma, J.D., Chanda, B.: Writer-independent off-line signature verification using surroundedness feature. *Pattern Recognition Letters* 33, 301–308 (2012)
14. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition:a comprehensive survey. *IEEE Trans.Pattern Anal. Mach. Intell.* 22(1), 63–84 (2000)
15. Seropian, A., Vincent, N.: Writers authentication and fractal compression. In: Eighth International Workshop on Frontiers in Handwriting Recognition, p. 434 (2002)
16. Sokal, R., Rohlf, F.: Biometry. W.H.Freeman and Co. (1969)
17. Prabhakar, S., Kittler, J., Maltoni, D., O'Gorman, L., Tan, T.: Introduction to the special issue on biometrics:progress and directions. *IEEE Trans.Pattern Anal.Mach.Intell.* 29(4), 513–516 (2007)

# Coordination of Electric Vehicles Charging to Maximize Economic Benefits

Yongwang Zhang<sup>1</sup>, Haoming Yu<sup>2</sup>, Chun Huang<sup>2</sup>, Wei Zhao<sup>1</sup>, and Min Luo<sup>1</sup>

<sup>1</sup> Electric Power Research Institute of Guangdong Power Grid Corporation,  
Guangzhou, China

<sup>2</sup> College of Electrical and Information Engineering, Hunan University,  
Changsha, China

**Abstract.** Under the constraints of distribution transformer capacity and customer charging needs, an coordinated charging model of electric vehicles is proposed to maximize the overall economic benefits of charging stations based on time periods of time-of-use(TOU) electricity price in power grids. Monte Carlo simulation method is utilized to generate the customer charging needs based on actual customer's charging profiles. The economic benefits of charging stations is simulated under uncoordinated and coordinated charging modes correspondingly. Simulation results have indicated that the economic benefits of the charging stations can be significantly improved by responding the TOU electricity price.

**Keywords:** Electric vehicles (EVs), economic benefits, time-of-use electricity price, Monte Carlo simulation, coordinated charging.

## 1 Introduction

The trend of global warming is increasing due to excessive emissions of greenhouse gases[1],[2]. As a new generation of transport, electric vehicles (EVs) have incomparable advantages with the conventional cars in its energy conservation and its reduction of the dependence on traditional fossil fuels. Currently, the development and application of EVs has been promoted by putting the appropriate policies around the world. It can be expected that, the charging of large-scale EVs hasn't negligibly impacted on the grid planning and operation in the future. One of the important impact is that the large-scale charging load for EVs will bring a new round of growth, especially the power load of the difference between peak and valley will further be exacerbated during the peak of EVs charging. It may lead to circuit overload of distribution network, voltage drop[3],[4], increasing losses of distribution network[5],[6], overload of distribution transformer[7],[8] and other issues.

On the other hand, as a new type of mobile load, the charging behavior of EVs has a strong spatial and temporal uncertainty. The difficulty of the grid operation and controlling is increased as large-scale charging of EVs. It is significantly important to reduce operational risk, improve efficiency and reliability of the power grid for the coordinated charging of EVs.

It is necessary to achieve the coordinated charging of EVs for charging stations. The ways of the coordinated charging is diverse. Eg, EVs will be seen as independent energy consumers, which is controlled by the charging control center of EVs in real time[9]. It can effectively reduce the loss of distribution system operation. The control method of coordinated charging is studied for reducing distribution losses based on the relationship among the network losses of distribution system feeder, the load rate of distribution network and the variance of load fluctuation[10]. On the basis of the stabilization of EV battery's life, the coordinated charging of EVs can reduce the charging costs of customers and provide ancillary service[11]. The use of the coordinated charging can reduce the negative impact on the grid caused uncertainty of new energy source output and EVs spatial and temporal profiles. A stochastic economic dispatch problem is studied to consider the uncertainty of EVs charging and wind-power output[12].

With the development of EVs, the coordinated charging of large-scale EVs has high demands about computing capabilities of the grid control center using centralized control while it is also facing challenges for the speed and reliability of real-time communication over a wide area of EVs in the control center. On the contrary, with only a relatively small amount of EVs, we can quickly collect charging information in real time, take into account the charging needs of customers and control coordinated charging of EVs based on real-time status of the grid in charging stations. On this basis, the coordinated charging of EVs will be able to be quickly and cost-effectively achieved combined with sub station and sub-district control.

This paper aims to study the coordinated charging of EVs in charging stations equipped with multiple charging piles and charging monitoring system. Under the constraints of distribution transformer capacity and customer charging needs, an coordinated charging model is proposed to maximize the overall economic benefits of charging stations based on time periods of TOU electricity price in power grids in charging stations.

## 2 Methodology

### 2.1 The Target of the Coordinated Charging

The conventional load and EVs charging load are connected with a distribution transformer. It can be considered that conventional load is zero when a charging station is only equipped with a exclusive distribution transformer.

The service providers of EVs charging achieve profitability by the difference between the charging tariffs and the purchased price from the grid companies in charging stations.

Whenever a EV can connect at No. n,(n=1,2,...,N) charger in charging stations, a coordinated charging control system(CCCS) may obtain EV's battery capacity  $B_n$ , as well as the current battery state of charge ( $SOC_n^A$ ) (the ratio of the current battery power divided by its total battery capacity). In order to develop the coordinated

charging strategy, the CCCS of charging stations needs to be informed the residence time  $t_n$  and the expected battery state of charge  $SOC_n^D$ .

## 2.2 Control Strategy of the Coordinated Charging

$N$  is the number of the charger in charging stations,  $P$  is the charging power of the charger, using the assumption of constant power charging in the charging process.  $S_T$  is the rated capacity of the distribution transformer,  $\lambda$  is the average power factor of the charging load.

According to the conventional load of the transformer history data, it can be forecasted that a conventional load curve for 96 points of day which the time interval is 15 minutes.  $A_j$  can represent the ratio of charging power for charging stations divided by transformer capacity with the  $j$ -th ( $j=1,2,\dots,96$ ) period of time daily.  $A_j$  can value  $[0, 1]$ . If the charging station is equipped with a exclusive distribution transformer,  $A_j \equiv 1$ .

The price information can include the purchased price from the grid companies and the charging price from customers that is  $c_j$  (yuan/kWh) and  $p_j$  (yuan/kWh) respectively.

The maximum time  $t_{\max}$  of all EVs residence time is determined from the current time, based on the set value of the current time and the expected residence time for all EVs in charging stations.  $J = \left[ \frac{t_{\max}}{15} \right]$  is the time period for coordinated charging control, and the CCCS can change a charge state every 15 minutes.

According to the time period  $J$ , a state matrix of charging stations  $S^{N \times J}$  is constructed, where  $S_{nj}$  can represent the parking status of NO.  $n$  charger at the  $j$ -th time.  $S_{nj} = 1$  means having a EV,  $S_{nj} = 0$  means EV-free.

Every 15 minutes, the CCCS for EVs can invoke the coordinated charging optimizer, calculate the on-off state each charger at the  $j$ -th time, thereby maximize the overall economic benefits in charging stations based on the information of the parking states, the customer needs, the grid load as well as electricity price.

## 2.3 Mathematical Optimization Model

To maximize the overall economic benefits of charging stations as the objective function, is given as

$$\max \sum_{j=1}^J \sum_{n=1}^N C_{nj} \times S_{nj} \times P \times \Delta t \times (p_j - c_j) \quad (1)$$

Where  $C^{N \times J}$  is the decision matrix of on-off state for all chargers in charging stations;  $C_{nj}$  is the control decision of NO.  $n$  charger at the  $j$ -th time,  $C_{nj} = 1$  can indicate that the charger is turned on,  $C_{nj} = 0$  can indicate that the charger is turned off;  $\Delta t$  can represent the length of the time period.

*Constraints:*

**a. Distribution transformer capacity constraints**

$$\sum_{n=1}^N C_{nj} \times S_{nj} \times P \leq A_j S_T \lambda, \forall j \in \{1, 2, 3, \dots, J\} \quad (2)$$

Where  $\lambda$  is the average power factor of the charging load.

**b. Charging demand constraints**

In the  $j$ -th time period, the battery SOC of the charging EV should at least reach the set value  $SOC_n^D$ , while the service providers should stop charging in the case of full SOC.

$$SOC_n^D B_n \leq (\sum_{j=1}^J C_{nj} \times S_{nj} \times P \times \Delta t + SOC_n^A \times B_n) \leq B_n, \forall n \in \{1, 2, 3, \dots, N\} \quad (3)$$

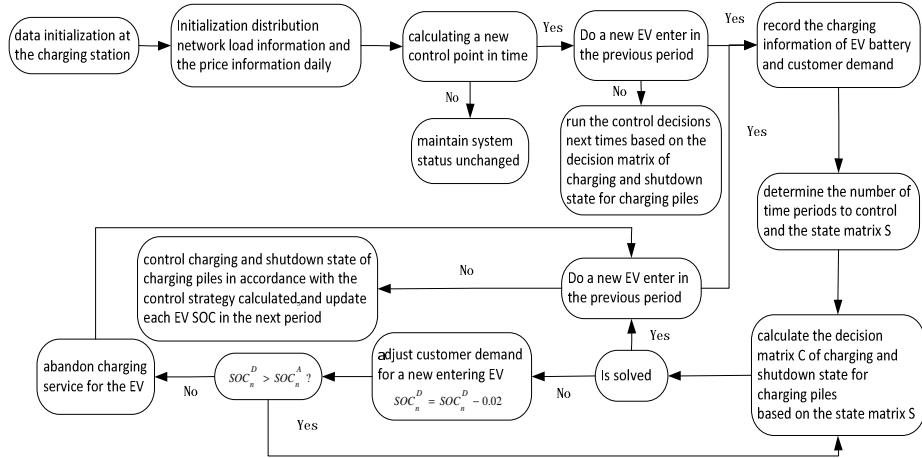
The above optimization model has a high efficiency using optimization tools that is linear integer programming model to solve.

## 2.4 Exception Handling

In addressing the actual needs of our customers, the service providers for EVs might encounter such a problem that the customer needs is urgent and requires to provide a lot of energy in a short time (such as larger  $SOC_n^D$ ,  $B_n$ , a small  $t_n$ ). Under constraints of the charging device hardware and transformer capacity, it can not be met that the final SOC at least reaches  $SOC_n^D$ . When solving the optimization problem, it shows no solution.

To solve this problem, the optimal control strategy is solved when the customer inputs  $SOC_n^D$ . If no solution, the CCCS can prompt the customer not to meet the charging demand, and make  $SOC_n^D$  decrement by 2% solving again until the problem have a solution. The CCCS can inform the customer the adjusted  $SOC_n^D$  ultimately. If the customer is satisfied with the adjusted  $SOC_n^D$ , the optimal control is implemented. If the CCCS can't make the customer satisfied and make  $SOC_n^D$  down to  $SOC_n^A$  which don't meet the charging needs in charging stations, the service providers can only give up the customer.

Based on the above model, the decision matrix C of on-off state for all chargers is proposed in the charging stations, achieving the coordinated charging control. After updating a state every 15 minutes, the CCCS may issue a new round of control command. If there is not a EV into the charging stations within 15 minutes, the CCCS can change the status of all chargers in accordance with the original computed control strategy. If there is a EV to enter into the charging stations, the status of all chargers is recalculated following the above steps. However, the SOC of the EVs is maintained constant in this period of 15 minutes. At the beginning of the next period time, the status of all chargers is changed based on the calculated control strategy in the charging stations. The process of coordinated charging control is shown in Figure 1.



**Fig. 1.** Block diagram of coordinated charging at the charging station

### 3 Modeling of Test Network

#### 3.1 Parameter Settings

In one residential charging station, for example, the conventional load and the charging load are connected with a distribution transformer that its capacity is 800kVA. The EVs are charged using the conventional charging mode that the charging power is 7kW and the power factor is 0.9. There is 80 chargers in the charging station. The curve of the proportion is residential load divided by distribution transformer capacity, where a maximum residential load is 50% of the distribution transformer capacity.

The charging service providers can adopt the form of industrial TOU electricity price purchased from the grid while taking a uniform price for EVs charging in charging stations. Table 1 shows the parameter settings of specific TOU electricity price.

**Table 1.** Parameter settings of energy prices in charging stations

	purchased price (yuan/kWh)	charging price (yuan/kWh)
valley time (0:00-8:00)	0.365	1
peak time (8:00-12:00, 17:00-21:00)	0.869	
common time (12:00-17:00, 21:00-24:00)	0.687	

Assuming that there is 100 private EVs for charging in the charging station every day, the general habits of customers driving EV is analyzed. Table 2 shows the charging data where  $N(a, b^2)$  can represent the normal distribution (mean=a, standard deviation=b),  $a \vee b$  can represent the bigger value between a and b,  $U(a, b)$  can represent a uniform distribution from the value of a to b.

**Table 2.** Charging parameters for EVs

Charging times/d	1	
initial charging time distribution(h)	$N(9, 0.5^2)$	$N(19, 3^2)$
charging probability for each period	0.2	0.8
expected charging time(h)	$U(0, 8)$	$U(6, 8)$
initial SOC distribution	$N(0.6, 0.1^2) \vee 0.2$	
EV battery capacity(kWh)	32	
EVs SOC when leaving	0.8 and 0.9 (probability of 0.5 respectively)	0.95 (arrive before 24h), else 0.9

### 3.2 The Uncoordinated Charging

In order to verify the effectiveness of coordinated charging mode, the operating conditions of the charging station and load conditions of the transformer are calculated under the uncoordinated charging mode, and the result of the conditions is compared with the coordinated charging mode.

Under the uncoordinated charging mode, the CCCS can provide continuous charging service for a new inserted EV as long as spare parking space in the charging station. The CCCS can't stop charging until the customer is leaving or this battery SOC for the EV has been fully charged. Therefore, due to a large number of EVs to access, the overload of the distribution transformer may be proposed. When the customer may has an urgent demand, the battery SOC may also can't be charged fully before leaving even if there has been charging.

### 3.3 Monte Carlo Simulation Method

Based on Monte Carlo simulation method, charging demand data of large-scale EVs daily is randomly generated, and the charging process is calculated under the coordinated and uncoordinated charging mode. It can analyze the impact on operating parameters of the coordinated charging in the charging station in Table 3.

**Table 3.** Simulation statistical information

Statistical Information	Statistical Methods
benefits of charging station daily	different between charging fees and buying electricity tariff at the charging station one day
the proportion of abandoning charging customers at the charging station daily	the ratio of giving up EVs customers divided by all EVs customers one day
the proportion of reducing the charging demand daily	the ratio of reducing the charging needs of customers divided by all EVs customers one day
the average time for computation	the mean of consuming calculation in 96 period one day
the percentage of minimum (maximum) load divided by distribution transformer capacity	the minimum(maximum)of load divided by distribution transformer capacity in the simulation

## 4 Results and Discussion

### 4.1 The Results of the Simulation

According to Monte Carlo method, the charging need for one hundred of EVs is simulated in a day. The CCCS can calculate economic benefits of the charging station daily, the proportion of the abandoned customers, the average proportion of the charging demand to reduce, the average calculating time and the percentage share of the maximum and minimum load divided by distribution transformer capacity under the un-coordinated and coordinated charging mode. The average earnings is simulated and calculated under the uncoordinated and coordinated charging mode.

As can be seen from the average yield curve, the average yield is still remained unchanged when Monte Carlo calculation times is more than 400. Therefore, the simulation times is setted up 400 times. Simulation is completed on the computer that CPU is Intel Core i3, 4G memory. The simulation results are shown in Table 4.

**Table 4.** Results of coordinated and uncoordinated charging modes

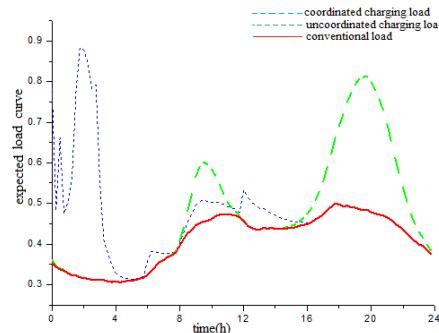
The charging control method	coordinated charging	uncoordinated charging
average benefits of charging station (yuan/d)	640.08	193.83
maximum benefits of charging station (yuan/d)	725.52	228.02
minimum benefits of charging station (yuan/d)	541.15	154.90
the average proportion of abandoning charging customers at the charging station daily(%)	2.22	1.67
the maximum proportion of abandoning charging customers at the charging station daily(%)	12	11
the minimum proportion of abandoning charging customers at the charging station daily(%)	0	0
the average proportion of reducing the charging demand daily(%)	8.01	\
the average time for computation(s/frequency)	1.1922	0.00015
the percentage of maximum load divided by distribution transformer capacity (%)	99.9	97.2
the percentage of minimum load divided by distribution transformer capacity (%)	30.6	30.6

Figure 2 shows two cases of expected EVs load curve under the uncoordinated and coordinated charging modes daily.

### 4.2 The Result of the Analysis

1) Under the coordinated charging mode, the overall economic benefits of charging stations is approximately three times than the uncoordinated charging mode.

This shows that the overall economic benefits are greatly increased by the introduction of coordinated charging method.



**Fig. 2.** Expected load curves under coordinated and uncoordinated charging modes

2) Compared with the simulation data, the ratio of the abandoned customers is kept at a low level under the two charging modes daily, indicating that the coordinated charging method does not give up more customers for charging service. The case of no entering into charging stations occurs substantially in no excess charging parking space. Therefore, the ratio of the abandoned customers is directly related to the number of parking space as well as the distribution of the charging time daily.

3) Under the coordinated charging mode, the average proportion of bringing down charging demand can remain at a very low level daily, indicating that it can meet the basic charging need having a higher economic efficiency in charging stations simultaneously.

4) Under the coordinated charging mode, the average time of control strategy for each time period is only about one second with a higher calculation speed. The algorithm is suitable for real-time and coordinated charging control in the large-scale of charging stations.

5) By analyzing the typical load curve of two scenarios daily, it can find that a large number of EVs access to charge in the evening peak load increasing further exacerbate difference between peak and valley. And under the coordinated charging mode, a lot of EVs may focus charging at the night valley to obtain a larger economic benefits due to a cheap purchased price although evening peak load didn't further increased. This may cause that there is a peak in the local power grid at night which is even higher than evening peak load. It can indicate that the CCCS can regulate the charging behavior in a simple TOU electricity price way making a lot of EVs gather in the period of cheap electricity price to charge. It can result in another peak to occur in the local power grid.

## 5 Conclusion

According to the real-time operating state, combined with the actual charging behavior, a model of the coordinated charging is proposed to maximize the overall economic benefits of charging stations achieving the coordination of charging control

considering the different SOC, the residence time and the customers' needs. The following conclusions are proposed through simulation analysis.

- 1) The proposed method of coordinated charging control can increase significantly the revenue of charging stations based on the customers' needs and transformer operation.
- 2) The model is proposed to solve mixed integer optimization suiting to large-scale charging stations, but depending on solving algorithm package of the control system installation-related problem.
- 3) From the load curve, we can see that only a single coordinated charging method based on the TOU electricity price may not reduce the difference between the peak and valley of the local power grid in certain circumstances. Otherwise, another peak load may occur in the local power grid owing to a large number of EVs to access.

**Acknowledgements.** The authors are grateful to the support of the National High Technology Research and Development Program ("863" Program) of China (No. 2012AA050211).

## References

1. Song, Y.H., Yang, X., Lu, Z.X.: Integration of plug-in hybrid and electric vehicles: experience from China. In: Proceedings of IEEE Power and Energy Society General Meeting, July 25-29 (2010)
2. Wen, C.K., Chen, J.C., Teng, J.H., et al.: Decentralized plug-in electric vehicle charging selection algorithm in power systems. *IEEE Trans. on Smart Grid* 3(4) (December 2012)
3. Singh, M., Kar, I., Kumar, P.: Influence of EV on grid power quality and optimizing the charging schedule to mitigate voltage imbalance and reduce power loss. In: Proceedings of Power Electronics and Motion Control Conference, September 6-8 (2010)
4. Olle, S., Carl, B.: Flexible charging optimization for electric vehicles considering distribution grid constraints. *IEEE Trans. on Smart Grid* 3(1) (March 2012)
5. Fernandez, L.P., San Roman, T.G., Cossent, R., et al.: Assessment of the impact of plug-in electric vehicles on distribution networks. *IEEE Trans. on Power Systems* 26(1), 206–213 (2011)
6. Acha, S., Green, T.C., Shah, N.: Effects of optimized plug-in hybrid vehicle charging strategies on electric distribution network losses. In: Proceedings of IEEE Transmission and Distribution Conference and Exposition, New Orleans, LA, USA, April 19-22, pp. 1–6 (2011)
7. Dow, L., Marshall, M., Le, X., et al.: A novel approach for evaluating the impact of electric vehicles on the power distribution system. In: Proceedings of IEEE Power and Energy Society General Meeting, Minneapolis, MN, USA, July 25-29, pp. 1–6 (2010)
8. Wu, D., Aliprantis, D.C., Ying, L.: Load scheduling and dispatch for aggregators of plug-in electric vehicles. *IEEE Trans. on Smart Grid* 3(1) (March 2012)
9. Clement-Nyns, K., Haesen, E., Driesen, J.: The Impact of charging plug-in hybrid electric vehicles on a residential distribution grid. *IEEE Trans. on Power Systems* 25(1), 371–380 (2010)
10. Sortomme, E., Hindi, M.M., MacPherson, S.D.J., et al.: Coordinated charging of plug-in hybrid electric vehicles to minimize distribution system losses. *IEEE Trans. on Smart Grid* 2(1), 198–205 (2011)

11. Rotering, N., Ilic, M.: Optimal charge control of plug-in hybrid electric vehicles in deregulated electricity markets. *IEEE Trans. on Power Systems* 26(3), 1021–1029 (2011)
12. Zhao, J., Wen, F., Xue, Y., et al.: Power system stochastic economic dispatch considering uncertain outputs from plug in electric vehicles and wind generators. *Automation of Electric Power Systems* 34(20), 22–29 (2010)

# Traffic Sign Recognition Using Perturbation Method

Linlin Huang and Fei Yin

School of Electronics and Information Engineering, Beijing Jiaotong University

No.3 Shangyuancun, Haidian, Beijing 100044, China

huangll@bjtu.edu.cn

National Laboratory of Pattern Recognition (NLPR)

Institute of Automation of Chinese Academy of Sciences, Beijing 100190, China

fyin@nlpr.ia.ac.cn

**Abstract.** Automatic traffic sign recognition (TSR) expects high accuracy and speed for real-time applications in intelligent transportation systems. Convolutional neural networks (CNNs) have yielded state-of-the-art performance on the public dataset GTSRB, but involve intensive computation. In this paper, we propose a traffic sign recognition method using computationally efficient feature extraction and classification techniques, and using the perturbation strategy to improve the accuracy. On the GTSRB dataset, using gradient direction histogram feature and learning vector quantization (LVQ) classifier achieves a test accuracy 98.48%. Using simple perturbation operations of image translation, the accuracy is improved to 98.88%. The accuracy is higher than that of single CNN and the speed is much higher.

**Keywords:** Traffic sign recognition, classification, perturbation.

## 1 Introduction

Automatic traffic sign recognition (TSR) from images and videos has important applications in intelligent transportation systems (especially, driver assistance). As a visual pattern recognition problem, it encounters difficulties due to cluttered image background, perspective distortion, lighting variation, low resolution, and sometimes, imprecise location. The solution of this problem relies on techniques of image pre-processing, feature extraction, classification and contexts utilization.

Aiming for application in driver assistance, many works have been accomplished on traffic sign detection and recognition from the 1990s [1][2]. The many existing recognition methods have utilized various feature extraction and classification techniques. The areas of pattern recognition and machine learning have offered a large selection of classifiers, all are applicable to TSR. The classification performance depends on the size and quality of training data, training algorithm and the feature representation of patterns. The classification methods that have been applied to TSR include cross-correlation [1], neural networks [2], support vector machines (SVMs) [3][4], hierarchical classifiers [5][6][7], similarity-based classifier [8], ensemble of binary classifiers with error-correcting output codes (ECOC) [9][10]. The feature

representation is more crucial to recognition performance because the separability between classes in the feature space largely depends on it. Many image features are available for TSR, include statistical and structural features, or global and local features. Recently, the histogram of oriented gradients (HOG) [11] is popularly and successfully used as a feature for image recognition, and has been applied to TSR [3][6][7]. Class-specific regional features have been used to improve the discriminability [12].

In recent years, deep neural networks (DNNs) have shown superiority in various image recognition tasks, including TSR. In the German Traffic Sign Recognition Benchmark (GTSRB) organized at the IJCNN 2011 [13][14], the best performance was yielded by a multi-column DNN (MCDNN) [15], and a multi-scale convolutional neural network (CNN) [16] performed competitively. DNNs, however, are very computationally expensive in both training and testing due to the large number of convolution and weighted combination operations. They are usually implemented in parallel computation with GPU, otherwise the computation is too slow. Wang et al. proposed a two-stage recognition method with SVM classification and color-based perspective adjustment in the second stage [7]. They obtained a higher accuracy than the MCDNN at much lower computation costs (test time 40ms per sample on a desktop computer).

In this paper, we propose an efficient method for TSR using perturbation with a smart baseline recognizer. The perturbation method has been widely used in image recognition to generate training samples for enhancing the generalization performance of classifiers, particularly in the methods based on DNNs. Perturbation on test samples has been shown effective in improving the test accuracy in character recognition [17][18], but has not been used in TSR, to our best of knowledge. Our baseline recognizer extracts from sign image multi-resolution gradient direction histogram (GDH) feature, which is similar to the HOG but is more accurate. The classifier is a nearest prototype classifier trained by a learning vector quantization (LVQ) algorithm. The baseline recognizer yields sufficiently high accuracy, say, 98.48% on the GTSRB dataset. Using simple perturbation operations of translation and rotation on the test sign image, the accuracy is promoted to 98.88%. The perturbation-based recognition process takes less 12ms in total on a test sample by single-core computation on CPU.

The rest of this paper is organized as follows. Section 2 introduces the overview of our method. Section 3 describes the details of our recognition method. Section 4 presents experiment results and section 5 concludes the paper.

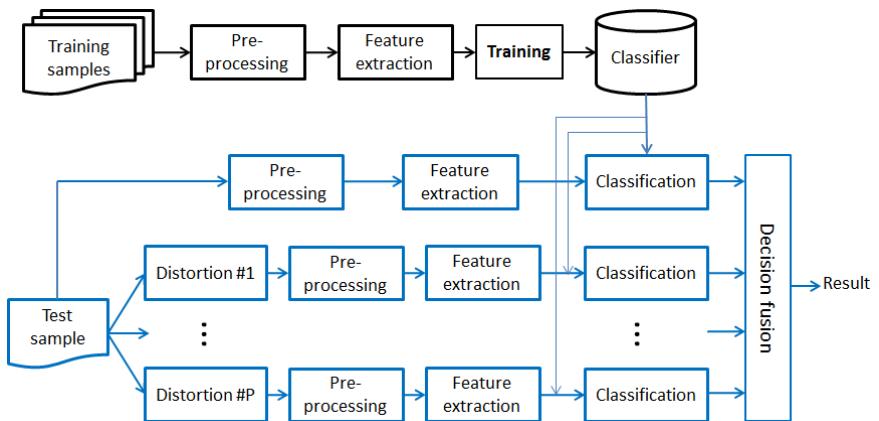
## 2 System Overview

In addition to the feature extraction and classification techniques, the availability of large set of training samples is important to the generalization performance of recognition. However, even a huge number of training samples cannot guarantee that the classifier can correctly recognize all variations. The perturbation method is taken to alleviate this problem in two ways. One way is to enhance the training dataset by

generating virtual samples by distorting real samples. Another way is to generate variations of test sample with the hope that at least one variation has regular shape and can be recognized correctly with high confidence. We tried both ways in our experiments and as the result, the first way did not improve substantially, while the second way resulted in significant improvement of test accuracy.

The overall diagram of the recognition system is shown in Fig. 1. Each sample (traffic sign image, either training sample or test sample) undergoes pre-processing and feature extraction to obtain a feature vector representation  $\mathbf{x} \in R^d$ . The classifier is trained with the feature vectors of the training samples. In ordinary pattern recognition, the test sample undergoes the same pre-processing and feature extraction procedures as training samples, and the obtained feature vector is classified by the trained classifier. By the perturbation method, the test sample undergoes a number of transformations (such as translation, scaling, rotation, shearing, perspective transformation, etc.) to generate a number of distorted samples, which undergoes pre-processing and feature extraction in the same way as the original test sample. The feature vectors are classified to output class labels and confidence scores, which are fused to give the final classification result.

The details of the pre-processing, feature extraction, classification, perturbation and decision fusion are given in Section 2.



**Fig. 1.** Overview of the training and recognition processes

### 3 Recognition Method

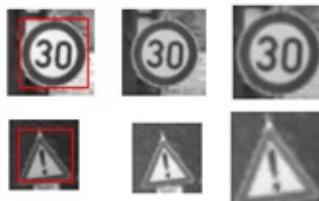
We choose efficient techniques to implement the baseline recognizer (including pre-processing, feature extraction and classifier) and perturbation. Pre-processing is to regulate the image size, position and gray intensity. For feature extraction, we adopt the gradient direction histogram (GDH) feature, which is similar to the HOG but is more accurate. For classification, we take the nearest prototype classifier. With few prototypes per class trained by LVQ-like discriminative learning, it yields fairly high classification accuracy at low computation complexity.

### 3.1 Pre-processing

Our recognition method does not utilize any color information. So, the sign image is first converted to gray-scale image by taking the average of RGB values. In the GTSRB dataset, each sign image has the boundary of sign area attached, and the size of sign area is variable. The gray intensity is highly variable depending on the illumination condition. We hence have two steps of image processing: intensity normalization and size normalization.

Intensity normalization is done by linearly transforming the gray levels of pixels in sign area such that the transformed values fit pre-specified mean and standard deviation (s.d.). We empirically set the mean and s.d. as 128 and 50, respectively.

For size normalization, the sign area is mapped to a standard image of given size (say, 40x40) by bilinear interpolation. For calculating gradient features, we enlarge the normalized image with an extra margin of two pixels, whose gray levels are interpolated from the input image. Fig. 2 shows some examples of gray intensity normalization and size normalization.



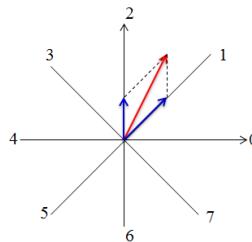
**Fig. 2.** Examples of gray intensity normalization and size normalization. In each row, left: original image; middle: intensity normalized; right: cropped and size normalized.

### 3.2 Feature Extraction

There have been many methods for feature extraction from images for shape recognition. The histogram of oriented gradients (HOG) [11] is widely used in recent years. In the area of character recognition, the histogram of gradient orientations/directions has been used popularly for feature extraction from the 1980s, and is still the state of the art. This class of feature is also called gradient direction histogram (GDH). We follow the GDH method of [19], which consists of two steps: gradient decomposition, Gaussian blurring.

Unlike that the HOG method quantizes the gradient direction into regions of angle, the GDH usually decompose the gradient vector into two sub-vectors using the parallelogram rule: the neighboring two standard directions are assigned gradient magnitude proportional to the length of the sub-vector (Fig. 3). The case in Fig. 3 has 8 standard directions but it is easily to generalize the rule to arbitrary number of standard directions. Specifically, for each pixel in the normalized image, we calculate the gradient using the Sobel operator, decompose the gradient into a number of standard directions. The gradient magnitude of each standard direction is stored in an image of same size as the normalized image, called direction map. From each

direction map, feature values are extracted by Gaussian blurring (low-pass filtering) and sub-sampling. The parameter of the Gaussian filter is determined according to the interval of sampling (size of block) [19]. The extracted values of different direction maps in the block form a local histogram. All the local histograms of different blocks or all the sampled values of different direction maps are concatenated to form a feature vector, whose dimensionality is  $D = n_d \cdot z_x \cdot z_y$ , where  $n_d$  is the number of standard directions, and  $z_x$  and  $z_y$  are the numbers of partitioned blocks in horizontal axis and vertical axis, respectively.



**Fig. 3.** Decomposition of gradient vector by parallelogram rule

Compared the HOG method, the GDH method has two advantages:

- Light computation. By parallelogram decomposition, the GDH method involves multiplication only and avoids calculating the angle of gradient, which is much more computation intensive than multiplication.
- Better translation invariance. By Gaussian blurring, the GDH feature is less variant to object translation. Though the HOG feature can alleviate the translation variance by extracting overlapping blocks, this largely increase the dimensionality of feature and complicates the subsequent classification.

### 3.3 Classification

The classification stage involves two techniques: dimensionality reduction and classifier. Dimensionality reduction can reduce the computation complexity of classifier and sometimes, improve the classification performance. We adopt the linear discriminant analysis (LDA) method for dimensionality reduction. The LDA maps the feature vector into a lower dimensional subspace spanned by the eigenvectors of largest eigenvalues of matrix  $S_w^{-1}S_b$ , where  $S_w$  is the within-class scatter matrix (average class covariance matrix) and  $S_b$  is the between-class scatter matrix, both estimated on the training dataset. Since the  $S_b$  has maximum rank  $M-1$  ( $M$  is the number of classes), the subspace has maximum dimensionality  $M-1$ . In the case of small number of classes, feature mapping to too low dimensional subspace will deteriorate the classification performance. The LDA incurs no loss of discriminability only when the classes all observe Gaussian densities with equal covariance. This is a strict assumption and usually does not hold in practice. To obtain higher dimensional

subspace, we adopt the regularized LDA method (called Fisher K-L method) of [20], which replaces  $S_b$  with  $S_b + \beta S_w$ , and  $S_w$  with  $(1 - \beta)S_w + \beta I$ , where  $\beta \in (0, 1)$  is a regularization parameter and  $I$  the identity matrix.

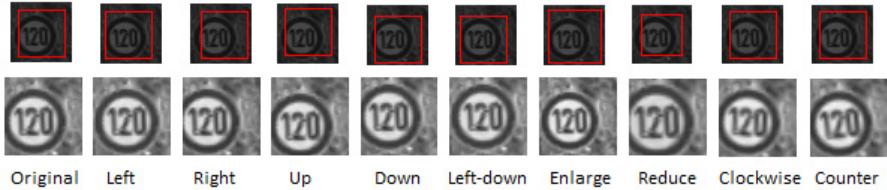
After dimensionality reduction, we use a nearest prototype classifier (NPC) for classification. It has very few prototypes per class, which are optimized using a LVQ-like discriminative learning algorithm. When using one prototype per class, the NPC is equivalent to a linear classifier under the Euclidean distance metric. After the LVQ algorithm of Kohonen [21], there have been many improved algorithms that optimize an empirical objective on the training data. We adopt the recent algorithm of logarithm of hypothesis margin (LOGM) [22]. In training, the feature vectors of training samples are fed iteratively. On each sample, the closest prototype of true class and the nearest one of rival classes are updated by stochastic gradient descent under an objective of logarithm loss.

### 3.4 Perturbation

The motivation of perturbation on test image is that compared to standard sign images, the test image is distorted in shape, translation, rotation, shearing, or other transforms. Translation is present because the sign is not located precisely. Either in sign detection or image processing after detection, the precise location of sign boundary is not trivial. Distorting the test sample may make it closer to the standard shape and improve the recognition accuracy. We design transformations of translation, scaling and rotation for perturbation. The formulations are as follows.

- Translation. We shift the boundary of sign to obtain translated sign image after normalization. In the original image, the boundary is moved to left, right, up, down, left-up, left-down, right-up and right-down. The distance of shift is empirically set as 1/40 of sign size (average of boundary width and height).
- Scaling. The boundary of sign is enlarged or reduced by two units of shift (one unit is 1/40 of sign size).
- Rotation. The sign image is rotated clockwise and counter-clockwise. The angle of rotation is empirically set as 3 degree.

In total, there are 8 perturbations of translation, 2 perturbations of scaling and 2 perturbations of rotation. Some examples of perturbations are shown in Fig. 4, where it can be seen that the normalized image with sign boundary shifted left-down is better standardized. Regarding the fusion of the classification results of multiple perturbations and the original image, we take the simplest way of selecting the result of highest confidence. In the case of nearest prototype classification, the confidence is reversely proportional to the nearest prototype distance. So, this is to select the result of minimum distance. With the hope that one perturbation of the original image is close to standard shape, the classification result of highest confidence is likely to be correct.



**Fig. 4.** Perturbations: five translated, two re-scaled, two rotated. Upper row: original image with sign boundary; lower row: normalized image. The rightmost two images are rotated ones.

## 4 Experimental Results

### 4.1 Dataset and Experimental Settings

We evaluated the performance of the proposed method on the GTSRB dataset and compare our results with those in the literature. The GTSRB dataset contains 51,839 sing images in total, partitioned into 39,209 training images and 12,630 test images. The signs are in 43 classes, including 8 speed limit signs, 4 other prohibitory signs, 4 derestriction signs, 8 mandatory signs, 15 danger signs, and 4 unique signs.

In our implementation of feature extraction, the sign area is normalized into size 32x32. The normalized image is partitioned into 8x8 blocks for extracting gradient direction histogram (GDH) feature. The gradient vector is decomposed into 12 standard directions. So, the dimensionality of gradient direction feature is 12x8x8=768D. We also extract gradient direction features from 16x16 normalized image, which is partitioned to 4x4 blocks, and 12x4x4=392D features are extracted. The features of two resolutions are concatenated into a 960D vector. For classification, the feature dimensionality is reduced by LDA or RLDA. By LDA, the subspace dimensionality was set as 42, while by RLDA, the subspace is 42D or 100D. The regularization parameter of RLDA was set as  $\beta = 0.1$ . In the nearest prototype classifier (NPC), the number of prototypes per class was set as 1, 2, and 3. In addition to the LVQ-LOGM algorithm for prototype learning, we also evaluated the NPC with prototypes as the cluster centers (by k-means clustering) of each class. The case of one mean prototype per class is also called as nearest mean classifier, which is combined with LDA dimensionality reduction in many works.

### 4.2 Results and Discussions

Table 1 shows the recognition accuracies on the test set of GTSRB dataset. In Table 1, GDH-768D is the feature of one resolution of normalized image (32x32), while GDH-960D is the feature of two resolutions. The number following LDA/RLDA indicates the subspace dimensionality.

First, on the single-resolution feature GDH-768D, we compare the performance of dimensionality reduction and classification methods. Comparing classifiers, it is apparent that discriminant prototype learning by LVQ-LOGM outperforms the k-means clustering-based prototype learning. In 42D subspace, the LDA and RLDA

perform comparably. When increasing the subspace dimensionality by RLDA, the test accuracy of k-means prototype classifier is decreased compared to 42D subspace, but the accuracy of LVQ-LOGM is increased significantly. This is because the higher-dimensional subspace preserves more discriminative information.

The two-resolution feature GDH-960D yields evident improvement of recognition accuracy compared to GDH-768D. The LVQ-LOGM prototype classifier with 3 prototypes per class yields the highest test accuracy 98.48%.

**Table 1.** Test accuracies (%) using GDH feature and nearest prototype classifier

	#prototype	k-means			LVQ-LOGM		
		1	2	3	1	2	3
GDH-768D	LDA-42	95.73	<b>97.28</b>	97.21	97.54	97.54	<b>97.66</b>
	RLDA-42	95.68	<b>97.35</b>	97.34	<b>97.67</b>	97.64	97.62
	RLDA-100	95.66	96.66	<b>96.67</b>	<b>98.04</b>	97.93	98.02
GDH-960D	LDA-42	96.48	97.48	<b>97.80</b>	97.93	97.98	<b>98.04</b>
	RLDA-42	96.34	<b>97.79</b>	<b>97.79</b>	97.94	<b>97.99</b>	97.95
	RLDA-100	96.33	97.04	<b>97.07</b>	98.04	98.35	<b>98.48</b>

We then applied perturbation to the best prototype classifier, LVQ-LOGM on GDH-960D. We tested three cases of perturbations: 8 perturbations of translation only, translation+scaling (10 perturbations), translation+scaling+rotation (12 perturbations). The test accuracies are shown in Table 2. Compare to the results without perturbation in Table 1, the test accuracy is improved significantly by perturbation. When using translation only, the accuracy of 3-prototype classifier is improved from 98.48% to 98.85. When using 12 perturbations including scaling and rotation, the accuracy is further improved to 98.88%. Nevertheless, the improvement of scaling and rotation is not so effective as that of translation only.

**Table 2.** Test accuracies (%) using perturbation with LVQ classifier

Perturbation	1-prototype	2-prototype	3-prototype
Translation	98.73	98.73	<b>98.85</b>
Transtration+scaling	98.75	98.79	<b>98.87</b>
Translation+scaling+rotation	98.76	98.80	<b>98.88</b>



**Fig. 5.** Sign images mis-recognized by the baseline recognizer but corrected by perturbation

Fig. 5 shows some sign images that were mis-recognized by the baseline recognition (GDH-960D and 3-prototype classifier) but was corrected by perturbation. There are also some cases that were correctly recognition by the baseline recognizer but were mis-recognized by perturbation (Fig.6), but overall, the correct rate is improved.



**Fig. 6.** Sign images that were correctly recognized by the baseline recognizer but mis-recognized by perturbation

Compared the results of German Traffic Sign Recognition Benchmark (GTSRB) in IJCNN 2011 [13][14], our result in only inferior to that of multi-column DNN (MCDNN, test accuracy 99.46%) , and is superior to the multi-scale CNN (98.31%). Even the result of our method without perturbation is superior to the multi-scale CNN. It is also interesting to compare the performance of GDH feature with that of the HOG. In [14], the HOG2 (1568D) reported the best result 95.68%, which is comparable to our result of single-resolution GDH using LDA classification, 95.73. Further, our GDH feature has much lower dimensionality 768D. When using two-resolution GDH (960D), the accuracy of LDA is improved to 96.48%. This indicates that the GDH feature is more accurate than the popularly used HOG.

Regarding the processing speed, our method of 3-prototype classifier on 960D GDH feature costs 0.91ms on a sign image on Intel Core i7-3770 CPU with programming in C++. When using perturbation of translation, the average processing time is 8.2ms, and when using perturbation of translation+scaling+rotation, it becomes 11.9ms. This is still much faster than the method of [7]. Note that the MCDNN of [15] has to be implemented in parallel computation.

## 5 Conclusion

In this paper, we proposed a traffic sign recognition method using perturbation. The baseline recognizer uses two-resolution gradient direction histogram classifier and discriminative prototype classifier. It yields fairly high accuracy of 98.48% on the GTSRB dataset at very high speed. By perturbation, the accuracy is improved to 98.88%. Our method is competitive in terms of the tradeoff between accuracy and speed compared to the state-of-the-art results in the literature. Our future work aims to further improve the accuracy using better perturbation operations and decision fusion strategy, and improve the processing speed of perturbation by hierarchical classification.

**Acknowledgments.** This work has been supported by the National Natural Science Foundation of China (NSFC) Grants 61271306 and 61175021.

## References

1. Piccioli, G., De Micheli, E., Paroli, P., Campani, M.: Robust method for road sign detection and recognition. *Image and Vision Computing* 14(3), 209–223 (1996)
2. de la Escalera, A., Moreno, L.E., Salichs, M.A., Armingol, J.M.: Road traffic sign detection and classification. *IEEE Trans. Industrial Electronics* 44(6), 848–859 (1997)
3. Greenhalgh, J., Mirmehdi, M.: Real-time detection and recognition of road traffic signs. *IEEE Trans. Intelligent Transportation Systems* 13(4), 1498–1506 (2012)
4. Maldonado-Bascon, S., Lafuente-Arroyo, S., Gil-Jimenez, P., Gomez-Moreno, H., Lopez-Ferreras, F.: Road-sign detection and recognition based on support vector machines. *IEEE Trans. Intelligent Transportation Systems* 8(2), 264–278 (2007)
5. Koncar, A., Janssen, H., Halgamuge, S.: Gabor wavelet similarity maps for optimising hierarchical road sign classifiers. *Pattern Recognition Letters* 28(2), 260–267 (2007)
6. Zaslavskiy, F., Stanciulescu, B.: Real-time traffic-sign recognition using tree classifiers. *IEEE Trans. Intelligent Transportation Systems* 13(4), 1507–1514 (2012)
7. Wang, G., Ren, G., Wu, Z., Zhao, Y., Jiang, L.: A hierarchical method for traffic sign classification with support vector machines. In: Proc. 2013 IJCNN, Dallas, USA (2013)
8. Paclik, P., Novovicova, J., Duin, R.P.W.: Building road-sign classifiers using a trainable similarity measure. *IEEE Trans. Intelligent Transportation Systems* 7(3), 309–321 (2006)
9. Baro, X., Escalera, S., Vitria, J., Pujol, O., Radeva, P.: Traffic sign recognition using evolutionary AdaBoost detection and forest-ECOC classification. *IEEE Trans. Intelligent Systems* 10(1), 113–126 (2009)
10. Escalera, S., Pujol, O., Radeva, P.: Traffic sign recognition system with beta-correction. *Machine Vision and Applications* 21(2), 99–111 (2010)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR, vol. 1, pp. 886–893 (2005)
12. Ruta, A., Li, Y., Liu, X.: Real-time traffic sign recognition from video by class-specific discriminative features. *Pattern Recognition* 43(1), 416–430 (2010)
13. German Traffic Sign Recognition Benchmark (GTSRB),  
<http://benchmark.ini.rub.de/>
14. Stallkamp, J., Schllipsing, M., Salmen, J., Igel, C.: Man vs. computers: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks* 32, 323–332 (2012)
15. Ciresan, D., Meier, U., Masci, J., Schmidhuber, J.: Multi-column deep neural network for traffic sign classification. *Neural Networks* 32, 333–338 (2012)
16. Sermanet, P., LeCun, Y.: Traffic sign recognition with multi-scale convolutional networks. In: Proc. 2011 IJCNN, pp. 2809–2813 (2011)
17. Yasuda, M., Yamamoto, K., Yamada, H.: Effect of the perturbed correlation method for optical character recognition. *Pattern Recognition* 30(8), 1315–1320 (1997)
18. Ha, T., Bunke, H.: Off-line handwritten numeral recognition by perturbation method. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(5), 535–539 (1997)
19. Liu, C.-L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten digit recognition: Benchmarking of state-of-the-art techniques. *Pattern Recognition* 36(10), 2271–2285 (2003)
20. Kimura, F.: On feature extraction for limited class problem. In: Proc. 13th ICPR, Vienn, vol. 2, pp. 191–194 (1996)
21. Kohonen, T.: Improved versions of learning vector quantization. In: Proc. IJCNN, vol. 1, pp. 545–550 (1990)
22. Jin, X.-B., Liu, C.-L., Hou, X.: Regularized margin-based conditional log-likelihood loss for prototype learning. *Pattern Recognition* 43(7), 2428–2438 (2010)

# A Novel Two-Stage Multi-objective Ant Colony Optimization Approach for Epistasis Learning

Pengjie Jing and Hongbin Shen

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University,  
and Key Laboratory of System Control and Information Processing,  
Ministry of Education of China, Shanghai, 200240, China  
`{jingse, hbshen}@sjtu.edu.cn`

**Abstract.** Recently, genome-wide association study (GWAS) which aims to discover genetic effects in phenotypic traits is a hot issue in genetic epidemiology. Epistasis known as genetic interaction is an important challenge in GWAS since it explains most individual susceptibility to complex diseases and it is difficult to detect due to its non-linearity. Here we present a novel two-stage method based on multi-objective ant colony optimization for epistasis learning. We conduct a lot of experiments on a wide range of simulated datasets and compare the outcome of our method with some other recent epistasis learning methods like AntEpiSeeker, Bayesian epistasis association mapping (BEAM) and BOolean Operation-based Screening and Testing (BOOST) method, finding that our method has a high power and is time efficient to learn epistatic interactions. We also do experiments in the real Late-onset Alzheimer's disease (LOAD) dataset and the results substantiate that our method has a potential in searching the suspicious epistasis in large scale real GWAS datasets.

**Keywords:** Multi-objective, Epistasis, Genome-wide association study, Single nucleotide polymorphism (SNP), Logistic regression, Bayesian network, Pareto optimal, Ant colony optimization, Chi-squared test.

## 1 Introduction

With the development of genome wide high-density single nucleotide polymorphisms (SNPs) genotyping technology, the genome-wide association studies (GWAS) are available and play a more and more important role in detecting the cause of disease [1]. Many diseases are influenced by multi-locus SNPs which mean they may have interaction with each other known as epistasis. Recently, a number of approaches have been proposed to detect epistatic interactions, including AntEpiSeeker [2], Bayesian epistasis association mapping (BEAM) [3], BOolean Operation-based Screening and Testing (BOOST) [4] and so on. Although these methods are said to be power and efficient according to their experiments, they suffer from low power, high false positives and inapplicable to high dimensional datasets.

In this paper, we propose a novel two-stage multi-objective method based on ant colony optimization algorithm. In the first stage, two objectives are combined to

model the association between phenotypes and genotypes. For the first one, standard logistic regression is used and the Akaike information criterion (AIC) score is designated as Objective 1. On the other hand, we adopt a Bayesian perspective to model the association and the K2 score is designated as Objective 2. The above two objectives are designed from the opposing schools of statistics, and our following results show they are complementary to each other and thus result in a better performance on general datasets. Then we optimize the two objectives based on the multi-objective ant colony optimization. In the second stage, we conduct exhaustive search of epistatic interactions with Chi-square test within the SNP subsets which are screened in the first stage. The experimental results on both simulated and real GWAS datasets substantiate that our method has a good performance and is scalable to high dimensional datasets.

## 2 Methods

Here are some notations used in this paper. Let  $X = \{x_1, \dots, x_m\}$  be a set of  $m$  SNPs. The  $x_i (1 \leq i \leq m)$  takes the value 0, 1, or 2, which corresponds to the homozygous major allele, heterozygous allele, and homozygous minor allele respectively. Suppose  $y$  denotes the disease status of individual, 1 for case and 0 for control. Let  $A_k = \{x_i, \dots, x_j\} (1 \leq i, j \leq m)$  denotes the  $k$ <sub>th</sub> SNP subset of  $X$ . The conditional probability of having the disease given the  $k$ <sub>th</sub> SNP subset is  $e = p(y=1|A_k)$ .

### 2.1 Logistic Regression

In statistics, logistic regression is a type of probabilistic statistical classification model. As discussed in [5], we assume a restricted model named additive interactive model (ADDINT) for epistasis learning. An example of ADDINT logistic regression model for two-locus interaction is in equation (1) which represents association relationship between SNPs  $x_1, x_2$  with the disease  $y$ .

$$\log \frac{e}{1-e} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \zeta x_1 x_2 \quad (1)$$

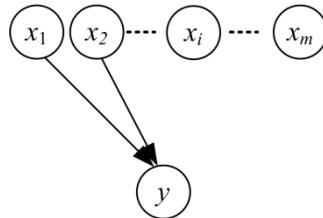
Where  $\beta_0$  denotes the mean term of model and  $\beta_1, \beta_2$  and  $\zeta$  are modeled as main effects and interactive effects, respectively. With logistic regression analysis used in the model, we can compute the log-likelihoods denoted as  $\log lik$  and the free parameters denoted as  $d$ . And then we can get the AIC score of the model as follows,

$$\text{AIC score} = -2\log lik + 2d \quad (2)$$

As the AIC score is designated as Objective 1, by comparing AIC scores with different SNPs in the model, the evidence for an effect on disease risk of different SNPs can be investigated. In this modeling approach, SNPs with low AIC score can be selected as such disease-correlated SNPs.

## 2.2 Bayesian Network

A Bayesian network (BN) is a probabilistic graphical model in which a set of random variables are denoted as a set of nodes and their conditional dependences are denoted as a set of edges in a directed acyclic graph (DAG). As in this particular GWAS problem, we can build a particular two-layer BN to represent the causative relation, where one layer consists of a set of SNP nodes and another of disease node. An example of the epistasis BN model in a set of SNP  $X$  can be seen in Fig 1.



**Fig. 1.** A 2-SNP epistasis BN model in a set of SNP  $X$

On the basis of the study in [6], in this paper we choose the K2 score derived from Bayesian scoring criteria to evaluate the BN model and take its log form:

$$K2 \text{ score}_{\log} = \sum_{i=1}^I \left( \sum_{b=1}^{r_i+1} \log(b) - \sum_{j=1}^J \sum_{d=1}^{r_{ij}} \log(d) \right) \quad (3)$$

Where  $I$  is the combinatorial number of SNP nodes with different values,  $J$  is the state number of disease node  $y$ ,  $r_i$  is the number of cases with SNP nodes take  $i_{\text{th}}$  combination,  $r_{ij}$  is the number of cases when the disease node takes the  $j_{\text{th}}$  state and his parents take the  $i_{\text{th}}$  combination. Thus we designate the K2 score as Objective 2, and the lower the log form score is, the stronger the association between the SNP subset and the disease is.

## 2.3 Pareto Optimal Approach

In the former sections, we introduce two different modeling approaches and their corresponding score functions. Then the problem of epistasis detecting is turned to be the problem of finding the best solution in respect of the two Objectives. For all solutions to both Objectives, we can classify them into two sets: a non-dominated set which comprises the optimal solutions to the multi-objective problem and a dominated set which should be neglected. Then the epistasis detecting problem can be extended to find a non-dominated set of solutions which is also known as “Pareto Optimal Set”. Here we use the non-dominated sort algorithm as follows to find the non-dominated set in the decision space  $A$  [7],

---

**Algorithm.** Non-dominated sort

---

**Input:** the SNP set  $A$ 

- 1:** for each  $A_i \in A$
- 2:**     mark( $A_i$ ) = non-dominated;
- 3:**     for each  $A_j \in A (i \neq j)$
- 4:**         if  $A_j < A_i$                       #If  $A_j$  dominates  $A_i$
- 5:**         mark( $A_i$ ) = dominated;
- 6:**         break;
- 7:**     end for
- 8:** end for

**Output:** all solutions with mark “non-dominated” are non-dominated set.

---

## 2.4 Ant Colony Optimization and Pearson’s $\chi^2$ test

To reduce the computational complexity of exhaustive searching for the non-dominated set, we introduce the ant colony optimization algorithm. The ants traverse routes and construct solutions which consist of any possible SNP combinations according to the pheromone values and transfer rules. Pheromone values are stored as a matrix  $\tau$ , and will reflect a signal distribution which is weak or strong according to the association between the corresponding SNPs and the disease. The transfer rules are described in equation (4) and (5) separately:

$$p_k(i, j) = \begin{cases} R & \text{if } (q \leq P_0) \\ 1 & \text{when } j = \text{rand}(U_k(i)) \quad \text{if } (q > P_0) \end{cases} \quad (4)$$

Where,  $p_k(i, j)$  is the probability according to which a ant selects SNP  $j$  followed SNP  $i$ ,  $q$  is a number generated randomly which is uniformly distributed in 0~1,  $P_0$  is a threshold to balance the convergence speed and avoidance of being trapped into local optimal solution.  $U_k(i)$  is the set of neighbor nodes of SNP node  $i$  that have not yet been visited by ant  $k$ , and  $\text{rand}(U_k(i))$  denotes the ant select a random SNP from the set  $U_k(i)$ .  $R$  represents the selection strategy of next added SNP by considering their pheromones’ distribution:

$$R = \begin{cases} \frac{\tau_{ij}^\delta \eta_j^\beta}{\sum_{u \in U_k(i)} \tau_{iu}^\delta \eta_u^\beta} & \text{if } (j \in U_k(i)) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

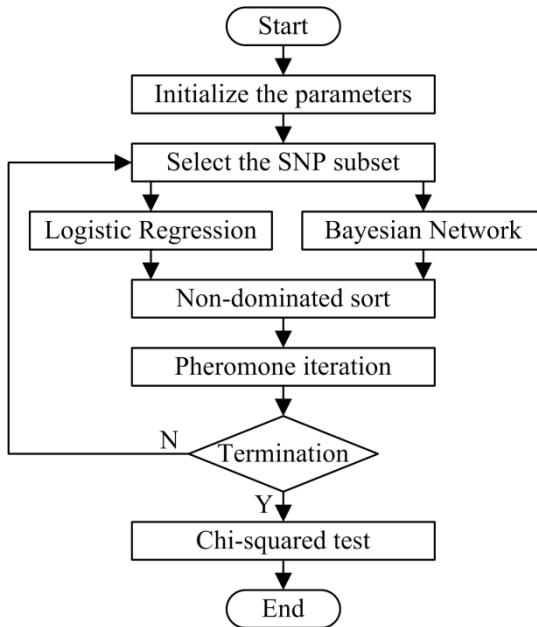
Where,  $\tau_{ij}$  is pheromone value between SNP  $i$  and SNP  $j$  and  $\eta_j$  is some form of priori information on SNP  $j$ ,  $\delta$  and  $\beta$  are parameters determining the weight of pheromone and the priori information on the SNPs respectively. In our work we let  $\eta$  and  $\delta$  equal to 1 so we treat each locus equally. The ACO algorithm gets the optimal solutions through a positive feedback which is formed by the pheromone iteration.

The pheromone value will be updated based on the performance of the SNP subset  $A_k$ , as:

$$\tau_{ij} \leftarrow (1 - \rho)\tau_{ij} + \rho\Delta\tau_{ij} \quad (6)$$

Where  $\rho$  is evaporate coefficient,  $\tau_{ij}$  is the pheromone value between SNP  $i$  and SNP  $j$ .  $\Delta\tau_{ij}$  is the changing pheromone value between SNP  $i$  and SNP  $j$ , which is equal to  $\lambda$  when  $A_k$  is belong to non-dominated set, otherwise it is equal to 0.

In the second stage, an exhaustive search of epistasis with Pearson's  $\chi^2$  test is conducted within the selected non-dominated solutions. Facing the problem of increased type I error in the presence of multiple testing, we implement a traditional and conservative Bonferroni correction. Thus the SNP subsets with  $P$ -values below the Bonferroni corrected significance level are reported by our algorithm. Fig 2 presents the flow chart of our algorithm:



**Fig. 2.** The flow chart of the algorithm

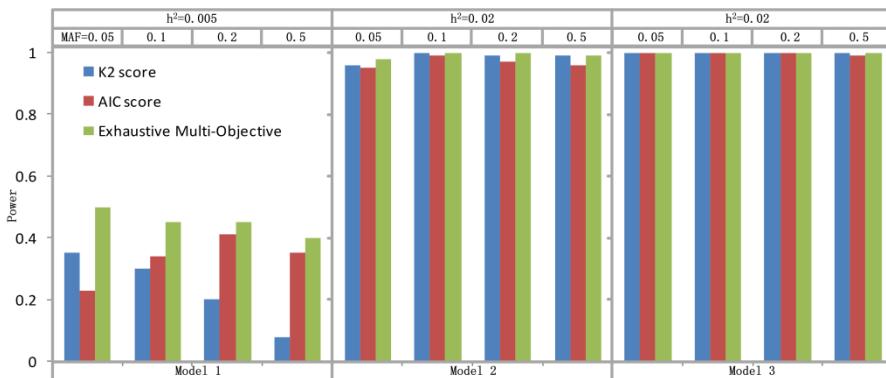
### 3 Experiments and Results

#### 3.1 Experiments on Simulated Datasets

**Epistasis Models.** An epistasis model is usually determined by three parameters: the disease prevalence  $P(D)$ , the genetic heritability  $h^2$  and the minor allele frequency  $MAF$ . In this paper, we consider three different classical epistasis models: Model 1 is a 2-locus multiplicative model; Model 2 is a 2-locus threshold model; Model 3 is a

2-locus concrete model used mimic the effect that epistasis has on susceptibility to handedness and the color of swine. In all three models, we set the *MAFs* of disease associated locus to be 0.05, 0.1, 0.2 and 0.5 to generate simulated datasets respectively and the *MAFs* of disease unassociated locus is obey the uniformly distribution of [0.05, 0.5]. To study the effect of the algorithm, 100 datasets are generated for each model with corresponding penetrance table and each dataset contains 100 SNPs and 1600 samples where each dataset contains equal number of cases and controls.

**Naive Multi-objective versus Single Objective.** Since the main scope of this study is to detect epistasis with multi-objective, we demonstrate there is an advantage of the combination of two objectives than single objective here. The power is designed as the ratio of the number of datasets in which the disease associated SNPs are successfully identified to all 100 datasets. Fig 3 presents the performance of comparison in 3 models and the results demonstrate multi-objective method outperforms each single objective method significantly. At the same time, the genetic heritability has more effect than the *MAF* in the ties between associated SNPs and disease as the detection power dramatically drop for low heritability while it has a moderate fluctuation for different *MAFs*.



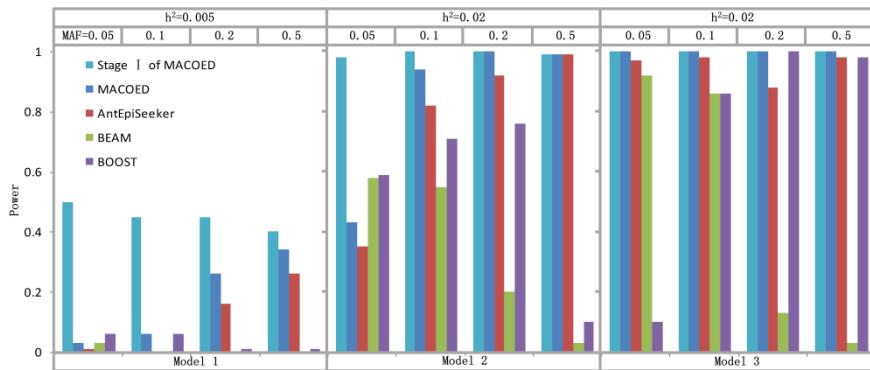
**Fig. 3.** Different power performances between single objective and multi-objective method

**Multi-objective Based on ACO versus Other Comparative Method.** Now we compare our algorithm with some commonly used algorithms including AntEpiSeeker, BEAM and BOOST. Here, we set the threshold *P*-value 0.1 after Bonferroni correction and other parameters according to previous studies [2].

We set the parameters of three comparative methods as their author recommended. In our method, we need to specify several parameters including  $P_0$ ,  $\rho$ ,  $\lambda$ , num\_ant (number of ants) and max\_iter (number of iterations). According to previous studies [2], a large  $P_0$ ,  $\rho$  and  $\lambda$  should be adopted for small number of SNPs in GWAS datasets (denoted as m) and small values for large m. The values of num\_ant and max\_iter are also determined by the m, where a large num\_ant and max\_iter should be adopted corresponding to a large m.

As the key idea of our method is multi-objective based on ACO algorithm, we output the intermediate non-dominated solutions in the screen stage and evaluate the non-dominated set by same designed power. Fig 4 presents the performances of comparison and shows the results of our algorithm have an increased power than other methods in most sets of parameters except some of the first model. This is because a tiny  $h^2$  and  $MAF$  may make Chi-square test lose its power in that the contingency table may have some empty cell. However, the power of intermediate results of our algorithm has significant superiority than other methods in all settings of 3 models denoting that the intermediate results are worth studying too.

In addition, our method is computationally efficient. In the above experiments, the total running times for all model of AntEpiSeeker, BEAM, BOOST and MACOED were 4.71, 5.04, 3.52 and 4.15 minutes respectively. The superior running time makes the GWAS in real large scale dataset available using our method.



**Fig. 4.** Different power performances among our algorithm and other comparative method

### 3.2 Real GWAS Dataset

Late-onset Alzheimer's disease (LOAD) is the most common form of Alzheimer's disease (AD) [8]. In GWAS filed, Rieman et al. found that 10 SNPs located in GAB2 gene on chromosome 11q14.1 are susceptible to LOAD in the presence of APOE  $\epsilon 4$  at the same time with a high statistical significance meaning that these GAB2 genes have an epistatic effect with APOE  $\epsilon 4$  in association with LOAD disease. After pre-processing, the LOAD dataset consists of 1,411 samples, and of them, 861 have LOAD and 550 do not. Each sample in this dataset consists of genotype information of 312,316 SNPs, APOE status and LOAD status.

Here we do 2-locus epistasis detection using our algorithm in SNPs of each separated chromosome combining APOE gene state. The 7 of our detected SNPs (rs4945261, rs2373115, rs1385600, rs7101429, rs10793294, rs1007837) are among these reported 10 SNPs located in GAB2. The remaining SNPs are scattered among chromosomes 1, 3, 4, 8 (2) and 18. Since there no biologically validated epistatic interactions for these remaining SNPs, they need further validation. The result shows that our method has practicality to detect epistasis in real GWAS dataset.

## 4 Conclusion

In this work, we provide a novel multi-objective method based on ant colony optimization algorithm to learn epistasis flexibly. We assess our method with three recent approaches by comparing their power and computational efficiency performances. Also we evaluate the availability of our method by comparing its performance with reported suspected gene locus on a real GWAS dataset, LOAD dataset. Our experimental results demonstrate that our method outperforms some other commonly-used methods in both simulated and real GWAS datasets.

In future work, we plan to find more powerful evaluation scores or appropriate efficient optimization strategies, and we can combine and embed them to our framework flexibly to increase its ability. Also we're going to do more testing works on other GWAS datasets to testify our framework's availability.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China (No. 61222306, 91130033, 61175024), Shanghai Science and Technology Commission (No. 11JC1404800), a Foundation for the Author of National Excellent Doctoral Dissertation of PR China (No. 201048), and Program for New Century Excellent Talents in University (NCET-11-0330).

## References

1. Churchill, G.A., et al.: The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* 36(11), 1133–1137 (2004)
2. Wang, Y., et al.: AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes* 3(1), 117 (2010)
3. Zhang, Y., Liu, J.S.: Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* 39(9), 1167–1173 (2007)
4. Wan, X., et al.: BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics* 87(3), 325–340 (2010)
5. North, B.V., Curtis, D., Sham, P.C.: Application of logistic regression to case-control association studies involving two causative loci. *Human Heredity* 59(2), 79–87 (2005)
6. Jiang, X., et al.: Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics* 12, 89 (2011)
7. Deb, K.: Multi-objective genetic algorithms: Problem difficulties and construction of test problems. *Evolutionary Computation* 7(3), 205–230 (1999)
8. Brookmeyer, R., Gray, S., Kawas, C.: Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *American Journal of Public Health* 88(9), 1337–1342 (1998)

# Hydraulic Excavators Recognition Based on Inverse "V" Feature of Mechanical Arm\*

Wenming Yang, Dedi Li, Daren Sun, and Qingmin Liao

Shenzhen Key Lab. of Information Sci & Tech / Shenzhen Engineering Lab.  
of IS & DRM

Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua  
University, China

**Abstract.** Detecting hydraulic excavators in videos can increase the confidence coefficient of illegal construction in nationalized land. Hydraulic Excavators have multifarious working postures making them a difficult target using even state of the art object recognition algorithms. The contribution of this paper is to propose an inverse "V" model for hydraulic excavator detection. In this paper, we describe an hydraulic excavator detection system based on inverse "V" feature of mechanical arm which is formed by boom and dipper and show a detection system. Then a real-time video processing method is presented which is used for monitoring illegal construction activities on a land of state-ownership.

**Keywords:** Computer vision, Object recognition, Hydraulic Excavator, Inverse "V" feature, Processing method for video.

## 1 Introduction

The urbanization process of China advance rapidly, therefore national land protection [1], especially, monitoring illegal construction on a land of state-ownership, becomes imperative mission for the relevant government. At present, the main regulation methods include remote sensing, human inspection and vehicle video monitoring. Poor real-time performance and lacking details exist in remote sensing [2]. Human inspection has the best flexibility, but it has the problem of low efficiency and high cost of labor resource. Although vehicle video monitoring save labor resources, it is unavailable for the bumpy areas. To overcome these problems, the intelligent video monitoring system is presented.

A variety of engineering vehicles such as hydraulic excavators, dump trucks and rollers are used in construction site. Among them, hydraulic excavators can handle multiple activities, such as excavation, loading, trimming and moving materials, so they are widely used in civil construction [3]. Detecting hydraulic excavator on a land of state-ownership can provide a reliable evidence for unauthorized construction.

---

\* This work was supported by the Science Industry Trade and Information Technology Commission of Shenzhen Municipality under Grant JCYJ20130402145002441.

In the following section, we firstly introduce the hydraulic excavator and the leading edge methods currently used to detect it, and then present our algorithm which was inspired by those methods for detecting the machine arm of the hydraulic excavator. Afterwards, the evaluation results of the algorithms on our database are presented. Finally, we give the detailed processing method for the online videos and make some discussion.

## 2 Hydraulic Excavator and Detection Methods

Hydraulic excavators are highly deformable machines. The mechanical arm which include boom, dipper and attachment can have countless forms, that means the machine can slew  $360^\circ$  and rotate all three parts of the arm around the hinged support. The typical deformations of the excavator are illustrated in Fig. 2. The reason of many scholars devote to detecting the mechanical arm rather than the whole excavator is that to detect excavator with limited number of training configurations as used in the case of rigid-frame equipment is impossible.

There are few literatures about detecting of excavators, but some articles which were related to analysis of production efficiency by excavators activities can be found in some articles. Zou etc. [3] analysis target segmentation by Hue, Saturation and Value in Color Space and then analysis its behavior. Robust detection algorithms such as Histogram of Oriented Gradients(HOG) and Harr-like features have been developed in the last decade. Azar etc. [4] divide the mechanical arm as root and part and use HOG as the feature, the method was set up to search for the root and then search for the part in the candidate regions. To improve the detection a spatial-temporal reasoning model is presented which restrict the moving pattern of a excavator by using time and space constraint.

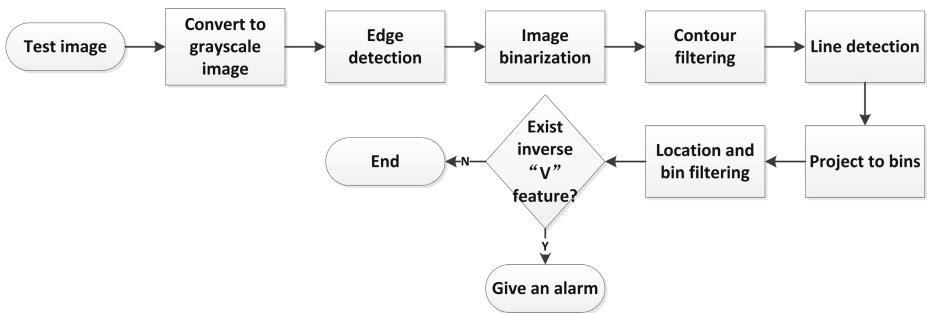
## 3 Detection Methods

### 3.1 Deformable Parts

An excavator can have countless forms in the process of operation, except that the boom is aligned with camera view, which is impossible to distinguish the parts [4]. Our goal is to detect the mechanical arm of the excavator. As illustrated in Fig. 3, the positions of boom and dipper [5] can be roughly divided into Horizontal, Left-inclined, Vertical, Right-inclined. "Horizontal" means the position of boom and dipper can roughly treated as horizontal. In other words, the "Horizontal" can be defined as  $-10^\circ$  to  $10^\circ$  and  $170^\circ$  to  $190^\circ$ . The benefit of such classification is to minimize the influence of the location of the excavator's body. We only need to pay attention to the location of boom and dipper, regardless of their hinged supports. The detection process is depicted in Fig. 1.

### 3.2 Diagonal line Detection

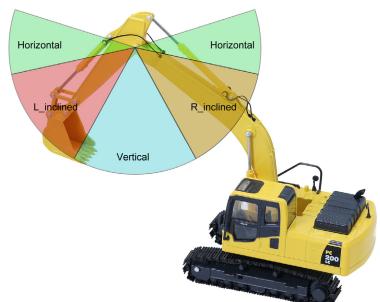
Considering the inverse "V" feature of mechanical arm, we employ the following Sobel mask which enhance the edge response on the direction of  $45^\circ$  and  $135^\circ$ .



**Fig. 1.** The flow chart of the detection algorithm



**Fig. 2.** Typical deformation of the hydraulic excavator



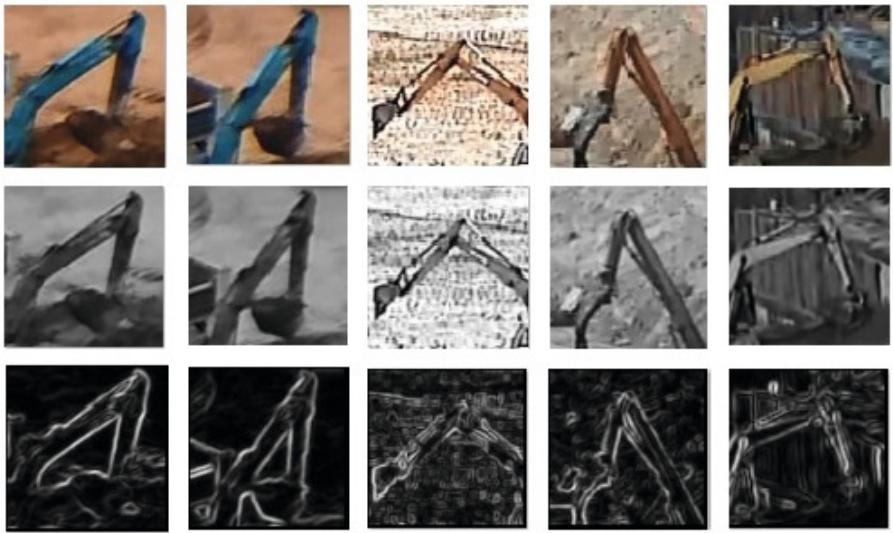
**Fig. 3.** The position of boom and dipper

We firstly convert the figure into grayscale, and then use the operators shown in Fig. 4 to detect the edge. The results are shown in Fig. 5.

2	1	0
1	0	-1
0	-1	-2

0	1	2
-1	0	1
-2	-1	0

**Fig. 4.** A pair of diagonal Sobel operators



**Fig. 5.** The first line is the original image. The second line is the grayscale image. The third line is the image processed by diagonal Sobel operators.

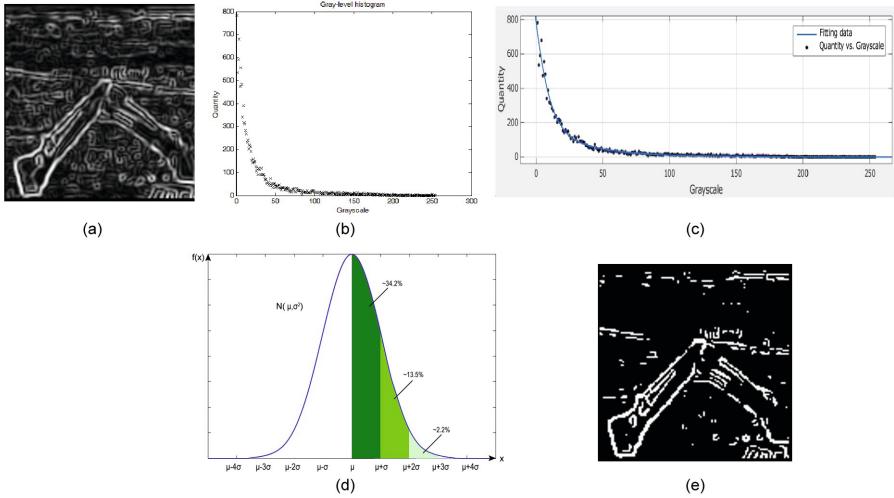
### 3.3 Using the Idea of Segmentation for Binary Operation

A grayscale image is obtained which indicates the diagonal edges. However, to acquire reliable edges, we need thresholding operation [6]. Global thresholding [7] and local thresholding [8] are two main thresholding algorithms which are used to get the binary image.

A robust method which can not only suppress the local noise but also keep the strong edges based on global threshold is needed. Inspired by the image segmentation [9], we can view background and noise as one class, and the strong edge as the other class. The gray histogram which can be treated as Gaussian distribution [10] is presented in Fig. 6.

The gray histogram can be denoted as Gaussian distribution according to the principle of energy concentration of Gaussian distribution which indicates between  $\mu - \sigma$  to  $\mu + \sigma$  gathering 68.3% of the energy and between  $\mu - 2\sigma$  to  $\mu + 2\sigma$  gathering 95.4 % of the energy [11].

In our algorithm, according to experience and experiment we choose  $\mu + 1.64\sigma$  as the threshold to get the binary image. The process is depicted in Fig.6. The binarization algorithm for gray scale image is described in Algorithm 1.



**Fig. 6.** (a) Sobel image (b) Gray-level histogram (c) Fitting data of the histogram (d) The energy distribution of Gaussian (e) Sobel-binary image

---

#### Algorithm 1. Binary operation based on the idea of segmentation.

---

- 1: Calculate the gray histogram of the image;
  - 2: Do Gaussian data fitting for the histogram;
  - 3: Get the threshold according to the energy of Gaussian distribution;
  - 4: Traverse each pixel in the graph and obtain the binary image based on the threshold;
- 

### 3.4 Contour Filter for the Binary Image

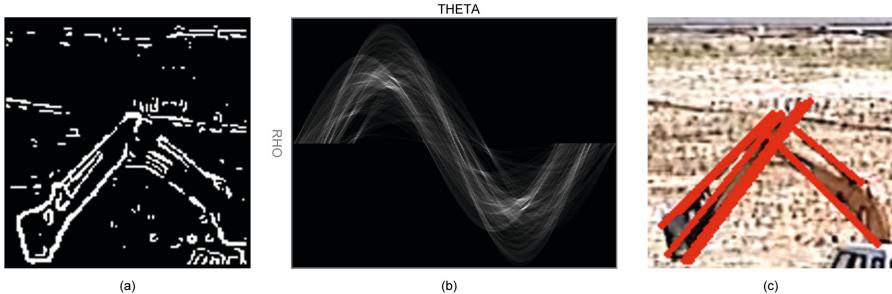
After the above operation, we generally obtain the diagonal edges. However, there still exist some annoying noises and pseudo edges. A straight idea is to use morphological operation to purify the image[12], but this is a challenging task under the complexity of the background, so morphological operations are unlikely to fulfill this job.

A contour filter [13] is adopted to get the result for the reason that the contour of the mechanical arm is larger than the rest and the other scattered contours are relatively small. In our project, we discard those contours whose perimeter is greater than 1000 or less than 400. The reason for that parameters selection is that it not only can get rid of the annoying small outlines but also can remove the big contours which were caused by data missing.

### 3.5 Line Detection

Hough transformation [14] is one of the classical methods for specific shapes detection. Compared with other methods, hough transformation can be better

at reducing the noise interference. Any point in the image space can be mapped to the transform domain as a line or sinusoid as illustrated in Fig.7. So a line can be determined by local extreme value in the transform domain [15]. Detailed algorithm is available in algorithm 2.



**Fig. 7.** (a) Binary image (b) Data distribution in Hough space (c) The result of lines detection

### 3.6 Bins and Positions Filtering

In this section, two filters are presented. The one which was mentioned in section 3.1 is used to divide lines into different types which represent as Horizontal, Left-inclined, Vertical, Right-inclined, and the other is to restrict the space position relationships of the lines.

---

#### Algorithm 2. Line detection based on Hough transform

---

- 1: **for** each pixel  $(x_0, y_0)$  **do**
  - 2:   Mapping a point into parameter space as a sinusoid;  
 $x_0\cos\theta + y_0\sin\theta = \rho;$
  - 3: **end for**
  - 4: Determine the accuracy of the parameter space;
  - 5: Determine local maximum values in the parameter space;
  - 6: Return lines according to the responding parameters;
- 

The process of the angle filter includes two steps. The first one is to divide the angle coordinate system into 4 bins, which represent as horizontal, right-inclined, vertical, left-inclined, respectively. However, to reduce false alarm we do not take horizontal bin into consideration in practical application. The second step is to project the lines into corresponding bins.

Location filter means the lines in corresponding bins should satisfy the corresponding position relationship. In details, left-inclined, vertical and right-inclined line's upper vertex should be close and the midpoints of the lines project to the ordinate should be roughly equal.

## 4 Experimental Result

An experiment was carried out to figure out if the method we proposed is valid. The experiment using 4271 images whose size 250\*150 pixels, containing 455 excavators varying in make and pose. These images were randomly grabbed from the construction videos taken in busy construction sites such that many of the images contain other types of the equipment [7]. Fig. 8 shown some detection samples.



**Fig. 8.** Detection samples

These images were detected with different thresholds, which alter the results. From Table 1, we find that the detection rate and false alarms reduces along with the increase of threshold. To have a better overview of the results, the receiver operating characteristic (ROC) curve is plotted and shown in Fig. 9.

## 5 Processing Method for Video

For a given video, the process of detection will be carried out as Fig.10. Firstly, we use 100 frames to construct the codebook model, and then detect whether there is motion object in the rest frames. If a block of foreground areas is detected, a recognition algorithm proposed above is carried out. In order to get better performance in the practical engineering, we adjust the threshold that have low detection rate and low false alarm. That will not affect the final result, for the reason that our aim is to detect hydraulic excavators in construction video and not to detect it for each frame. The method achieves promising results in online video system.

**Table 1.** Results of the method

		Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9
455 positive samples	True Positive	417	307	232	141	90	78	67	56	46
	False Negative	38	148	223	314	365	377	3888	399	409
3816 negative samples	True Negative	3801	3801	3809	3811	3811	3811	3813	3813	1815
	False Positive	15	13	7	5	5	5	3	3	1

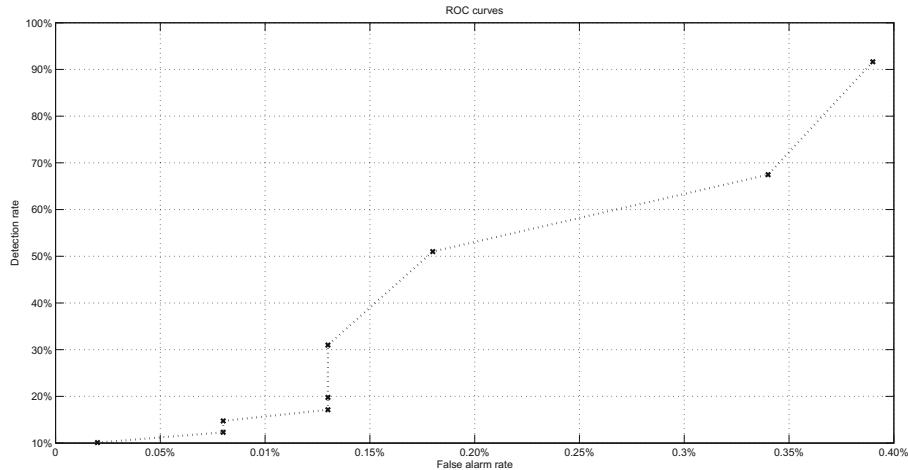


Fig. 9. ROC curve of the results on the test dataset

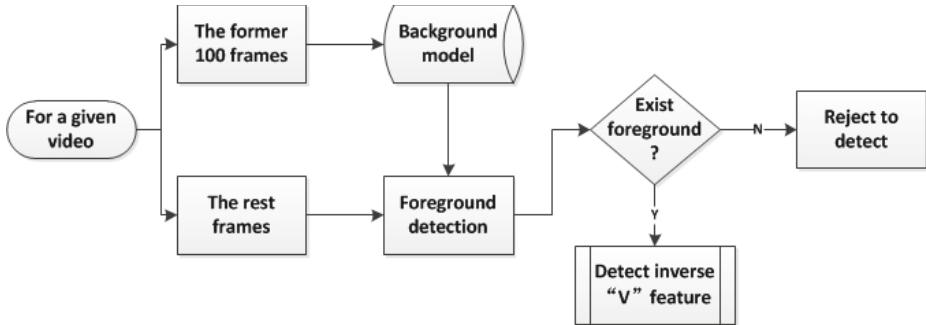


Fig. 10. The flow chart for video process

## 6 Conclusion

Inverse "V" feature model of excavator's machine arm is introduced to detect hydraulic excavator in this paper. This method first detect reliable edges in the image and then project them to the corresponding bins, and give the final result according to the position relationship. In order to improve the efficiency and accuracy, we first get the motion areas and then search mechanical arm's Inverse "V" feature in the motion areas. This combination showed promising result in recognizing excavator in online videos from stationary cameras. Future work will be focus on spatial-temporal reasoning [16] to eliminate false alarm and other identification methods.

## References

1. Skinner, M.W., Kuhn, R.G., Joseph, A.E.: Agricultural land protection in china: a case study of local governance in zhejiang province. *Land Use Policy* 18(4), 329–340 (2001)
2. El Amrani, C., Rochon, G.L., El-Ghazawi, T., Altay, G., Rachidi, T.-E.: Development of a real-time urban remote sensing initiative in the mediterranean region for early warning and mitigation of disasters. In: 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2782–2785. IEEE (2012)
3. Zou, J., Kim, H.: Using hue, saturation, and value color space for hydraulic excavator idle time analysis. *Journal of Computing in Civil Engineering* 21(4), 238–246 (2007)
4. Azar, E.R., McCabe, B.: Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in Construction* 24, 194–202 (2012)
5. Mcleod, C.C.: Excavating machine. US Patent 2,452,632 (November 2, 1948)
6. Weszka, J.S.: A survey of threshold selection techniques. *Computer Graphics and Image Processing* 7(2), 259–265 (1978)
7. Al-amri, S.S., Kalyankar, N.V., et al.: Image segmentation by using threshold techniques. *arXiv preprint arXiv:1005.4020* (2010)
8. Ahmad, M.B., Choi, T.-S.: Local threshold and boolean function based edge detection. *IEEE Transactions on Consumer Electronics* 45(3), 674–679 (1999)
9. Pal, N.R., Pal, S.K.: A review on image segmentation techniques. *Pattern Recognition* 26(9), 1277–1294 (1993)
10. Goodman, N.R.: Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *Annals of Mathematical Statistics*, 152–177 (1963)
11. Rasmussen, C.E.: Gaussian processes for machine learning (2006)
12. Wang, D., Haese-Coat, V., Ronsin, J.: Shape decomposition and representation using a recursive morphological operation. *Pattern Recognition* 28(11), 1783–1792 (1995)
13. Catanzaro, B., Su, B.-Y., Sundaram, N., Lee, Y., Murphy, M., Keutzer, K.: Efficient, high-quality image contour detection. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2381–2388. IEEE (2009)
14. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM* 15(1), 11–15 (1972)
15. Canny, J.: A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 679–698 (1986)
16. Teal, M.K., Ellis, T.J.: Spatial-temporal reasoning based on object motion. In: BMVC, pp. 1–10 (1996)

# Real-Time Traffic Sign Detection via Color Probability Model and Integral Channel Features

Yi Yang and Fuchao Wu

Institute of Automation, Chinese Academy of Sciences, Beijing, China  
[{yangyi,fcwu}@nlpr.ia.ac.cn](mailto:{yangyi,fcwu}@nlpr.ia.ac.cn)

**Abstract.** This paper aims to deal with real-time traffic sign detection. To this end, a two-stage method is proposed to reduce the processing time with little influence to AUC (area under curve) value. In first stage, a color probability model is proposed to transform an input image to probability maps. The traffic sign proposals are then extracted by finding maximally stable extremal regions on these maps. In second stage, an integral channel features detector is employed to remove false positives of the proposals. Experiments on the GTSDB benchmark [1] show that the proposed color probability model achieves the highest recall rate and the proposed two-stage method significantly improves computational efficiency with good AUC value in comparison with the state-of-the-art methods.

**Keywords:** Color Probability Model, MSER, Integral Channel Features, Traffic Sign Detection, Real-Time.

## 1 Introduction

Traffic sign detection plays an important role in Driver Assistant System and Intelligent Autonomous Vehicles. The rich information contained by traffic signs helps to maintain the traffic order and improve safety. There are many difficulties in traffic sign detection, such as illumination changes, color deterioration, motion blur, cluttered background and partial occlusion. These difficulties make traffic sign detection still an unsolved problem.

As traffic signs appear clearly in the road environment due to their well-defined colors and shapes, most traditional methods of traffic sign detection focus on utilizing color and shape information. As summarized in [2,3], the most popular color-based method is pixel-wise color threshold segmentation. However, it is difficult to select a perfect threshold to segment all images. To avoid this problem, [4] proposes a novel Eigen-Color model to detect the pixels with higher reflectance from background, [5] proposes a probabilistic measure for color pre-processing, and [6] regards color segmentation as a classification problem by using an SVM classifier. In shape-based methods, Hough transform [3], [7] and radial symmetry detector [3], [8] are most widely used. A method based on the

vertex and bisector transformation is proposed in [9] for triangular sign detection. The fast Fourier transform (FFT) of shape signature is adopted in [10]. [11] detects traffic signs by finding maximally stable extremal regions (MSERs [12]) from gray image for traffic signs with white background and from normalized red/blue image for traffic signs with red or blue backgrounds. [6] employs shape matching by convoluting the image with manually designed shape templates after color processing to find traffic sign candidates. As shape is not sensitive to lighting changes but could be easily affected by cluttered background, several methods propose to combine both color and shape information to improve detection performance as described in [3].

In this paper, we propose a two-stage method for real-time traffic sign detection. Firstly, a new color probability model is proposed to compute the probability maps, in which the high intensities indicate the presence of the specific colors (red and blue) of traffic signs. Next, MSERs are extracted on these probability maps and a non-maximum suppression procedure is performed to discard repeated MSERs. The remaining MSERs are regarded as traffic sign proposals. Then, an integral channel features detector [13,14] is applied on these proposals to produce the final detected traffic signs. Note that we only apply the integral channel features detector to traffic sign proposals rather than using the traditional multi-scale sliding window strategy. This would greatly save the processing time with little influence to AUC value. The overall workflow of our method is illustrated in Fig. 1.

The remainder of this paper is organized as follows. Section 2 describes the proposed method, including the color probability model and integral channel features detector. Experiments are reported in section 3. Section 4 concludes this paper.

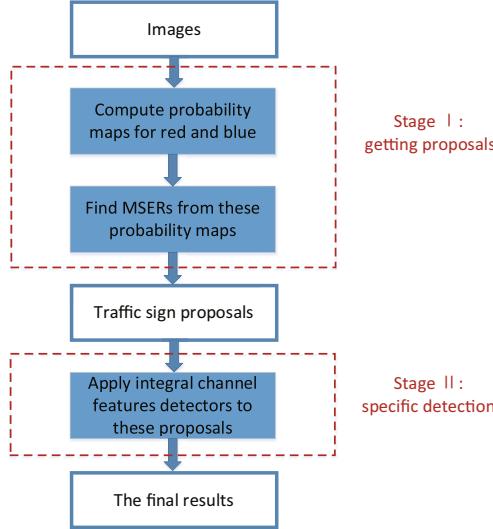
## 2 Our Method

### 2.1 Color Probability Model

The color of traffic signs is quite distinct from that of background, making it an intuitive feature for traffic sign detection. As the original RGB values are easily affected by various lighting conditions, they are often normalized or converted to HSV values to improve robustness [2].

In this section, we propose a new color probability model to compute the probability of belonging to a color of traffic sign for each pixel in image. Such a model is built based on the real distribution of traffic sign colors estimated from manually collected training samples. To improve robustness to lighting changes, we also convert RGB values to Ohta space [15] as it performs best in our experiments.

Assume there are  $N - 1$  colors for traffic signs, and all the backgrounds are denoted by another one color. Firstly, We manually collect the RGB values of these  $N$  colors from training images. Next these RGB values are converted to



**Fig. 1.** The overall workflow of the proposed method

Ohta space by

$$\begin{aligned} P_1 &= \frac{1}{\sqrt{2}} \frac{R - B}{R + G + B} \\ P_2 &= \frac{1}{\sqrt{6}} \frac{2G - R - B}{R + G + B} \end{aligned} \quad (1)$$

where  $P_1$  and  $P_2$  are the normalized components presented in [16]. Then, the mean vector  $\mu_i$  and covariance matrix  $\Sigma_i$  of color  $i$  in Ohta space are estimated from the collected samples.

Let  $x = (P_1, P_2)$  be the normalized components of Ohta space for a pixel, and  $C_i$  denotes the class of  $i$ -th color. According to Bayesian rules,  $P(C_i|x)$  is calculated by

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}, \quad i = 1, \dots, N \quad (2)$$

As  $P(x)$  denotes the color probability of the pixel  $x$ , it is a constant in a specific image. Therefore, the calculation of  $P(C_i|x)$  can be simplified as

$$P(C_i|x) = P(x|C_i)P(C_i), \quad i = 1, \dots, N \quad (3)$$

where  $P(x|C_i)$  and  $P(C_i)$  are the likelihood and prior, respectively. We simply use Gaussian distribution to model the distribution of a color class in Ohta space. Our experiments show that the Gaussian distribution performs quite well. Thus,  $P(x|C_i)$  can be computed by

$$P(x|C_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu_i)\Sigma_i^{-1}(x - \mu_i)^T\right\}, \quad i = 1, \dots, N \quad (4)$$

where  $D$  is the dimension of  $x$ . The calculation of prior is straightforward. We estimate the prior  $P(C_i)$  by the ratio of the number of samples in  $C_i$  and the total number of samples in all  $N$  classes:

$$P(C_i) = \frac{\#C_i}{\sum_{k=1}^N \#C_k}, \quad i = 1, \dots, N \quad (5)$$

where  $\#C_i$  is the number of samples in color class  $C_i$ . We further normalize  $P(C_i|x)$  to  $[0, 1]$  by

$$P'(C_i|x) = \frac{P(C_i|x)}{\sum_{k=1}^N P(C_k|x)}, \quad i = 1, \dots, N \quad (6)$$

By computing  $P'(C_i|x)$  for all color classes of traffic signs, we obtain  $N - 1$  probability maps.

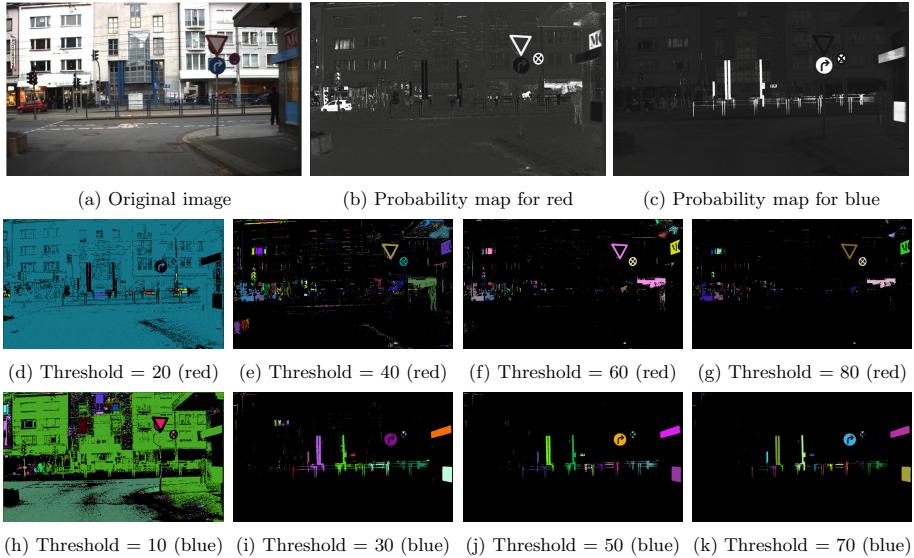
As the high intensities in probability maps indicate the presence of color of traffic signs, we further employ an MSER region detector [12] to find stable regions as traffic sign proposals. Fig. 2 shows an example of the first stage. The top row shows the original image and two probability maps computed by color probability model: (a) original image, (b) the probability map for red and (c) the probability map for blue. The red pixels in (a) have high intensities in (b) and the blue ones have high intensities in (c). The middle and bottom rows of Fig. 2 illustrate the connected components by thresholding the probability maps at four different levels. We denote different connected components with different colors for clarity. To find MSERs, value of each pixel in the probability map is set to 1 if its intensity is larger than a threshold, otherwise 0. Then the most stable connected components which maintain their shapes while thresholding the probability map at several levels are extracted as MSERs, i.e. the traffic sign proposals. As shown in Fig. 2, the red triangle and blue circle in the original image maintain their shapes in (e)-(g) and (i)-(k), respectively. Finally, a standard non-maximum suppression is performed to remove repeated proposals.

**Fast Computation.** As the transformation from RGB space to Ohta space and the calculation of probability are quite time consuming, the color probability model can not be directly used in real-time traffic sign detection.

To deal with this problem, a pre-calculated look up table (LUT) is used to speed up the computation. More specifically, we firstly compute the probabilities  $P'(C_i|x)$  and store them in the LUT offline. This would make an LUT with  $256^3 \times N$  elements. During the online detection, we simply compute the index of each pixel by its RGB values and find its corresponding probability in the LUT. With the use of LUT, the time for computing probability maps for a  $1360 \times 800$  image could be reduced from several minutes to about 30 ms on our PC (Intel 4-core 3.1GHz CPU, 4G RAM).

## 2.2 Integral Channel Features Detector

To achieve a high detection rate, the first stage should keep a high recall rate. This would make the traffic sign proposals contain lots of false positives. To filter

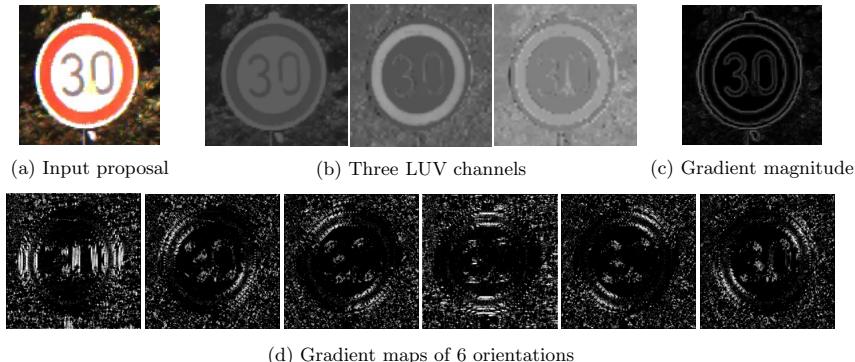


**Fig. 2.** Top row: Original image and the probability maps for red and blue. Middle and bottom rows: Connected components while thresholding the probability maps at some threshold levels.

out these proposals, we use an integral channel features detector in the second stage.

The general idea of integral channel features is to compute features on various channels of the input image by using integral image. A channel is a registered gray image of the input image, which typically includes intensity, color, integral histogram, gradient histogram, linear filters (eg. Gabor filters or Difference-of-Gaussian filters) and nonlinear filters (eg. gradient magnitude or Canny edges). The features are randomly generated rather than carefully designed by hand-craft. We only consider first-order feature which is a sum over a rectangular region in a given channel [13]. A large pool of candidate features are firstly generated by randomly chosen channel indexes and rectangles. Then the selected features are learned by Boosting and ‘soft cascade’ [17].

In our method, we use the combination of LUV color, gradient histogram and gradient magnitude as the channel features since such a combination presents the best performance in [13]. An example of the selected channels is shown in Fig. 3. We generate 1960 candidate features from these channels for each traffic sign proposal and train the boosted classifier with 16384 weak classifiers. Note that we only perform the integral channel features detector on traffic sign proposals rather than using the traditional multi-scale sliding window strategy. This would greatly save the processing time.



**Fig. 3.** An example of the selected channels of traffic sign proposal. (a): Input traffic sign proposal. (b): Three color channels of LUV space. (c): Gradient magnitude. (d): Gradient maps of 6 orientations.

### 3 Experiments

In this section, we carry out our evaluation on the German Traffic Sign Detection Benchmark (GTSDB) [1] which is the first and unique standard data set for traffic sign detection. In GTSDB, there are a total of 900 images (600 for training and 300 for testing) with the size of  $1360 \times 800$ . The traffic signs are divided into three categories, including Prohibitory signs with red color and circular shape, Danger signs with red color and triangular shape, and Mandatory signs with blue color and circular shape. Some examples of traffic signs in GTSDB are shown in Fig. 4. We manually collect color samples from the first 200 training images to build the color probability model, and train the integral channel features detectors for the three categories from all the 600 training images.

To show the effectiveness of the proposed color probability model (CPM), we firstly compare it with two other methods which also utilize color information: RGBN+gray [11] and SVM [6]. As the outputs of all these methods are gray images, we use MSER region detector with the same parameter setting to find traffic sign proposals. Due to the difference in the output gray images, the MSER region detector may extract different number of traffic sign proposals. For each traffic sign category, we test 15 parameter settings, and the results are shown in Fig. 5. The x-coordinate denotes the average number of traffic sign proposals on 300 testing images for each proposal set. The y-coordinate presents the recall rate of each proposal set. It can be found that the proposed CPM obtains the highest recall rate with the least number of proposals in all three categories. Such a superior performance demonstrates that the proposed CPM could well distinguish the color of traffic signs from the background. Due to the high recall rate of CPM, the number of inputs of the integral channel features detector can be greatly reduced. Furthermore, we conduct an experiment to show the overall performance by taking the proposals produced by these three methods as the



(a) Traffic sign examples of the three categories

(b) Some images in GTSDB

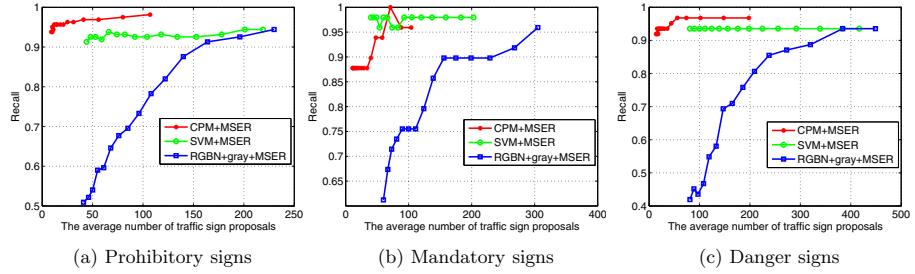
**Fig. 4.** (a) Some traffic sign examples of the three categories. Top row: Prohibitory signs. Middle row: Danger signs. Bottom row: Mandatory signs. (b) Some images in GTSDB.

input of integral channel features detectors (ICFD). The precision-recall curves are shown in Fig. 6. It is clear that the detector achieves the best performance by using CPM.

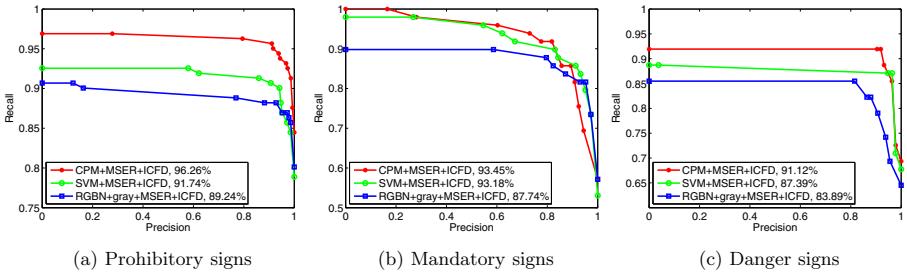
**Table 1.** The AUC values and processing times of the three methods

	Prohibitory	Mandatory	Danger	Time (seconds)
Our method	97.46%	<b>93.45%</b>	91.12%	<b>0.3</b>
Method in [6]	<b>100%</b>	92%	<b>98.85%</b>	0.4-1
ICFD in [14]	97.12%	93.34%	93.45%	0.6

We also compare the performance of our overall framework with the stat-of-the-art method in [6] and the original integral channel features detector in [14]. The AUC value is used as the evaluation criterion. The results are shown in Table 1. Our method performs the best in Mandatory signs while [6] performs the best in Prohibitory signs and Danger signs. Note that, the performance of our method is comparable to the performance of the original integral channel



**Fig. 5.** The recall rates and the corresponding average number of proposals of the three methods



**Fig. 6.** The precision-recall curves of the three methods. For each category, the parameters of MSER region detector and the integral channel features detectors are the same. The AUC values are also listed in the figure.

features detector on Prohibitory signs and Mandatory signs and only a little worse on Danger signs. However, our method obtains notable improvement in speed. For a  $1360 \times 800$  image, [6] takes 0.4-1 second on a PC with an Intel 4-core 3.7GHz CPU and 8G RAM. This makes it unsuitable for a real-time application. In contrast, our method only takes about 0.3 second on a PC with Intel 4-core 3.1GHz CPU and 4G RAM. In addition, the integral channel features detector is performed on the full image in [14] by using a modified sliding window strategy, which replaces the densely sampled image pyramid by a combination of sparsely sampled image pyramid and classifier pyramid, to reduce detection speed. However, it is performed on traffic sign proposals in our method. The processing time of the two methods proves that it is more efficiency with the stage of getting proposals than performing sliding window on full image directly.

## 4 Conclusion

In this paper we present a new framework for real-time traffic sign detection. The framework includes two stages: the first stage extracts traffic sign proposals by using a new color probability model and an MSER region detector. The second stage filters out the false positives of the proposals by employing an integral

channel features detector. Experiments on GTSDB benchmark show that the color probability model could extract proposals with a high recall rate and the two-stage method could significantly improve the computational efficiency with good AUC values. This makes the proposed framework a good choice for real-time traffic sign detection. Furthermore, some other strategies could be used to further improve efficiency, such as the methods proposed in [18,19].

**Acknowledgement.** This work is supported by the National Science Foundation of China (91120012).

## References

1. Houben, S., Stallkamp, J., Salmen, J., et al.: Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In: International Joint Conference on Neural Networks, pp. 1288.1–1288.8. IEEE press, Dallas (2013)
2. Gmez-Moreno, H., Maldonado-Bascn, S., Gil-Jimnez, P., et al.: Goal evaluation of segmentation algorithms for traffic sign recognition. *IEEE Transactions on Intelligent Transportation Systems* 11(4), 917–930 (2010)
3. Mogelmose, A., Trivedi, M.M., Moeslund, T.B.: Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems* 13(4), 1484–1497 (2012)
4. Tsai, L.W., Hsieh, J.W., Chuang, C.H., et al.: Road sign detection using eigen colour. *IET Computer Vision* 2(3), 164–177 (2008)
5. Houben, S.: A single target voting scheme for traffic sign detection. In: Intelligent Vehicles Symposium, pp. 124–129. IEEE press, Baden-Baden (2011)
6. Liang, M., Yuan, M., Hu, X., et al.: Traffic sign detection by ROI extraction and histogram features-based recognition. In: International Joint Conference on Neural Networks, pp. 1483.1–1483.8. IEEE press, Dallas (2013)
7. Garcia-Garrido, M.A., Sotelo, M.A., Martm-Gorostiza, E.: Fast traffic sign detection and recognition under changing lighting conditions. In: Intelligent Transportation Systems Conference, pp. 811–816. IEEE press, Toronto (2006)
8. Ruta, A., Li, Y., Liu, X.: Real-time traffic sign recognition from video by class-specific discriminative features. *Pattern Recognition* 43(1), 416–430 (2010)
9. Belaroussi, R., Tarel, J.P.: Angle vertex and bisector geometric model for triangular road sign detection. In: Workshop on Applications of Computer Vision, pp. 1–7. IEEE press, Snowbird (2009)
10. Gil-Jimnez, P., Bascn, S.M., Moreno, H.G., et al.: Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies. *Signal Processing* 88(12), 2943–2955 (2008)
11. Greenhalgh, J., Mirmehdi, M.: Real-time detection and recognition of road traffic signs. *IEEE Transactions on Intelligent Transportation Systems* 13(4), 1498–1506 (2012)
12. Matas, J., Chum, O., Urban, M., et al.: Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing* 22(10), 761–767 (2004)
13. Dollr, P., Tu, Z., Perona, P., et al.: Integral Channel Features. In: British Machine Vision Conference, London, pp. 91.1–91.11 (2009)
14. Dollr, P., Belongie, S., Perona, P.: The Fastest Pedestrian Detector in the West. In: British Machine Vision Conference, Aberystwyth, pp. 68.1–68.11 (2010)

15. Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. *Computer Graphics and Image Processing* 13(3), 222–241 (1980)
16. Vertan, C., Boujemaa, N.: Color texture classification by normalized color space representation. In: 15th International Conference on Pattern Recognition, pp. 580–583. IEEE press, Barcelona (2000)
17. Zhang, C., Viola, P.A.: Multiple-Instance Pruning For Learning Efficient Cascade Detectors. In: Neural Information Processing Systems, Vancouver, pp. 1681–1688 (2007)
18. Pang, Y., Yuan, Y., Li, X., et al.: Efficient HOG human detection. *Signal Processing* 91(4), 773–781 (2011)
19. Pan, J., Pang, Y., Zhang, K., et al.: Energy-saving object detection by efficiently rejecting a set of neighboring sub-images. *Signal Processing* 93(8), 2205–2211 (2013)

# Study of Charging Station Short-Term Load Forecast Based on Wavelet Neural Networks for Electric Buses

Lei Zhang, Chun Huang, and Haoming Yu

College of Electrical and Information Engineering, Hunan University, Changsha, China

**Abstract.** With the large-scale use of electric vehicles (EVs), a short-term load forecast method based on wavelet neural network (WNN) for electric buses is proposed to analyze load characteristics in order to better arrange transmission and distribution planning and regulate EVs charging or discharging, which comes from the current measured data related the charging station, Guangdong. This method is used to predicting EVs' load data of two test day selected randomly, compared with the effect of the single BP network model. The statistical results show that the prediction method has higher accuracy to meet certain application requirements than BP network applying to short-term load forecast of charging station for electric buses.

**Keywords:** Electric buses, wavelet neural network, short-term load forecast, BP network, charging station.

## 1 Introduction

Because of the excellent ability of function approximation, artificial neural network (ANN) as an important method to power system short-term load forecasting has found more and more recognition and application. As the drawbacks like determining difficultly the number of hidden nodes, the ANN has been improved by many other algorithms such as fuzzy theory, expert systems and genetic algorithm, which made great superiority in power load forecast especially in the short-term power load prediction [1].

Wavelet neural network (WNN) is a feed-forward network, which is a novel algorithm combined by wavelet theory and ANN. Nowadays, WNN has been widely used for many fields, such as troubleshooting, sample classification, function approximation, load forecast, and so on[2].

In this paper, a novel model is established to predict the EV charging load and count example validation based on the historical EV charging load. Although the forecast results of EV charging load may be a little deviation with the actual amount, this method is mainly used to estimate the future large-scale charging load levels in order to provide the basis for the impact on the grid, charging and discharging control strategy, charging infrastructure planning. If the development large-scale EV has become a reality, then short-term load forecast has a big significance to the EV charging load as part of the power system load. The electric grid requires arranging reasonable schedules and supply plans based on the EV charging demand. Under the

coordinated charging and V2G scenario, it can arrange to implement controlled policies and TOU to optimize the large-scale EVs' coordinated charging and discharging according to the results of the charging demand forecast, in order to mitigate the effects of the charging load on the grid, to reduce the cost of charging, to shift load peak and fill valley, to provide support services and other targets [3][4][5].

## 2 WNN Methodology

### 2.1 WNN Summarize

WNN is a neural network model based on wavelet analysis and neural network theory. According to the using of hidden nodes parameters, WNN can be broadly divided into two categories: one is the discrete WNN which values discrete hidden nodes parameters, and the discrete time-frequency analysis method is generally designed to determine the number of hidden nodes and network parameters easily for function approximation and signal analysis; the other is continuous WNN which values continuous hidden nodes' parameters, It can constitute a group of small wavelet radix from a mother wavelet scaling and translation, then you can use a linear combination of wavelet radix to approximate all of the functions in arbitrary precision with the condition that a sufficient number of hidden nodes.

Either continuous or discrete WNN, the linear combination of the wavelet base must be dense in  $L^2(R)$ . This is the theoretical guarantee that WNN has a good performance [6].

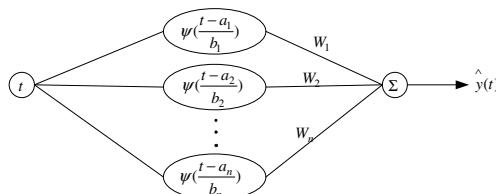
Compared with conventional ANN, WNN has the following advantages:

a. WNN can use wavelet theory to determine the network parameters and the number of hidden nodes, so it could avoid some problems like redundancy and misconvergence in traditional methods.

b. Because of wavelet bases is orthogonal, the solution is existence and uniqueness. With linear weighting coefficients and convex learning objective, compared with traditional algorithms based on BP neural networks, this method can well solve the convergence rate and local minimum trouble of it.

c. The parameters of WNN have clear physical meaning in time and frequency field because the wavelet scaling factor corresponds frequency and the shift factor corresponds time.

### 2.2 Continuous WNN with Single Input and Single Output



**Fig. 1.** Continuous WNN with single input and single output

The output of continuous WNN is

$$\hat{y}(t) = \sum_{n=1}^N W_n \psi\left(\frac{t-b_n}{a_n}\right) + \theta \quad (1)$$

It needs more hidden nodes with using sigmoid function to fit a complicated function, and that produces considerable errors. By comparison, continuous wavelet transform has higher prediction accuracy with the same number of hidden nodes due to its time-frequency localization characteristics.

Now we make detailed comparison among the properties of wavelet neural networks and BP networks.

BP neural network usually take radial basic function as exciting function of a hidden layer, which is specified as follows:

$$f(x) = 1/(1+e^{-x}) \quad (2)$$

The n-th input neuron of hidden layers of BP neural network is

$$net_n = W_n t + u_n \quad (3)$$

The output of BP neural network is

$$\hat{y}(t) = \sum_{n=1}^N W_n f(net_n) + \theta = \sum_{n=1}^N W_n f(W_n t + u_n) + \theta = \sum_{n=1}^N W_n f\left(\frac{t - (-u_n/W_n)}{1/W_n}\right) + \theta \quad (4)$$

The continuous WNN has the same network structures with BP neural network which have the same number of network parameters (including weights and thresholds). And WNN differs from the BP neural network only in exciting function it uses, which use the wavelet function instead. The adjustment of hidden layer weights and thresholds in BP neural network has been replaced by the adjustment of scale parameters and translation parameters in WNN. Compared to BP neural network, WNN has reliable a theoretical basis and be not easy to lost in local minimum in training process. WNN is suitable to predict for fluctuation signal and has higher accuracy than BP neural network with the same number of neural network nodes [7]. Continuous WNN has a similar structure, less hidden nodes and stronger anti-interference ability than BP neural network [2].

WNN with single input and single output can be easily extended to that with multiple inputs multiple outputs, which can consider the impact of various factors on the load of electric buses. So it has greater flexibility and applicability.

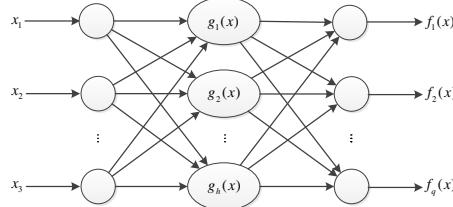
### 3 Load Forecast of Charging Station for Electric Buses

#### 3.1 The Characteristics Analysis of Short-Term Load Based on WNN

As an important part of the energy management system (EMS), power system load forecasting can guarantee electric system safely, stably and economically, and it's also the basis of grid scheduling plan and power supply plan to grid company.

Considering the load volatility of charging station for the electric buses and many other influenced factors, a prediction load model based on WNN is proposed.

Because of the development of EV is small and few data about charging load, there are few electric vehicle charging load forecasting work at home and abroad[15].With analyzing historical charging load data of a bus station in Guangzhou, the short-term load forecasting model has been established based on WNN.



**Fig. 2.** Three layers WNN network structure

The load forecast model of charging station for electric buses by WNN network consists of three layers. The time unit of load forecasting is 15 minutes, 96 points one day.

With the same type of the charging load, the input is maximum temperature, minimum temperature, the type of weather and date; the output is charging load which we need to predict. The nodes of the input is  $n = 100$ , the nodes of the output is  $m = 48$ . According to the literature [14], we could get the number of the hidden layer nodes:

$$N = \sqrt{nm + 1.6799n + 0.9298} \quad (5)$$

Morlet wavelet is selected as wavelet function of the hidden layer, which has good time-frequency locality, symmetry and clear expressions of time domain. Since the wavelet parameters are continuous, the number of wavelet functions is infinite and relevant, so the wavelet coefficients have a great redundancy. WNN is formed based on the accuracy requirements by optional part of the wavelet function, with less the number of hidden nodes, better anti-jamming performance and relatively stable numerical calculation than BP neural network. Its expression is:

$$g(x) = e^{-\frac{x^2}{2}} \times \cos(1.75x) \quad (6)$$

The network was trained with historical electric power load data of 35 days, and the result of the short-term electric power load forecast is successful.

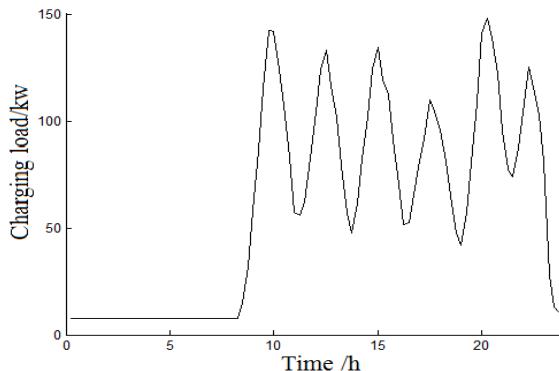
### **3.2 The Charging Load Processing and Analysis of Electric Buses Charging Station**

The station has 14 charging piles and 14 electric buses. It collected electric bus charging load data by SCADA from October, 2013 to July, 2014. Every 15 minutes is a load record collection points, and 96 points a day.

To extract principal features of the signal is the key to predict electric bus load accurately. According to the application experience of neural network, the solution can be more accurate by more hidden nodes. But that causes the generalization capabilities decreasing of neural network.

Because of many factors such as electromagnetic interference, human error and so on, there are many wrong historical charging load data. So, the data is modified with two methods in this paper: one is that if a charging load is 0, it replace with the mean load of its adjacent two; the other is that if one charging load exceed the mean of all charging loads, it replaced with the mean.

The average charging load curve of the charging station for electric buses in Guangdong province is shown in Figure 3.



**Fig. 3.** The average charging load curve of the charging station for electric buses

There are buses to begin charging in charging station around 8:00 every day, and there are generally six charging peak with buses running strong regularity.

#### 4 Examples of Analysis to Predict

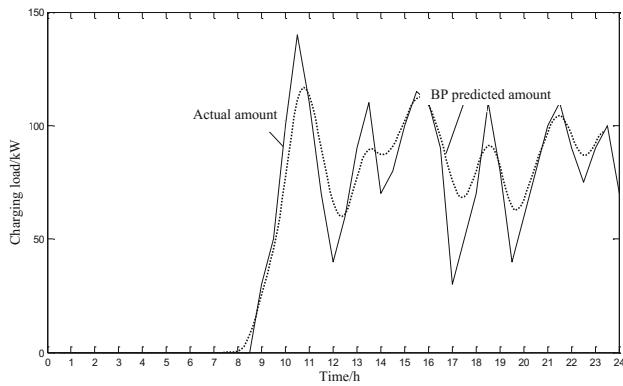
Four records are randomly selected in a electric bus charging stations of Guangdong province for the simulation, which is in June 16, 2013 (Monday), June 21, 2013 (Saturday), December 16, 2013 (Monday) and December 21, 2013 (Saturday).

BP neural network and WNN are used to research and process the records separately with the same structure. The network consists of three layers, and the input, hidden and output layers have 24, 24 and 24 neurons respectively, and realized the approximation of a continuous function.

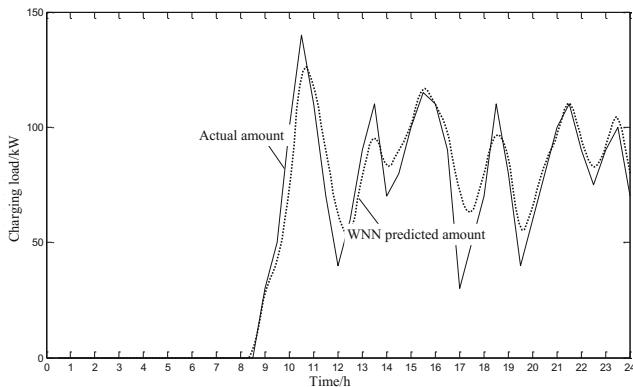
The forecast results are shown in Figure 4 to Figure 7 using BP neural network and WNN.

From Fig 4 to Fig 7 we can see that the results by WNN are closer to the actual amount than that by BP neural network, and the lines calculated by WNN are smoother than that by BP neural network. Then we use two indicators to show that the WNN is better than BP neural network.

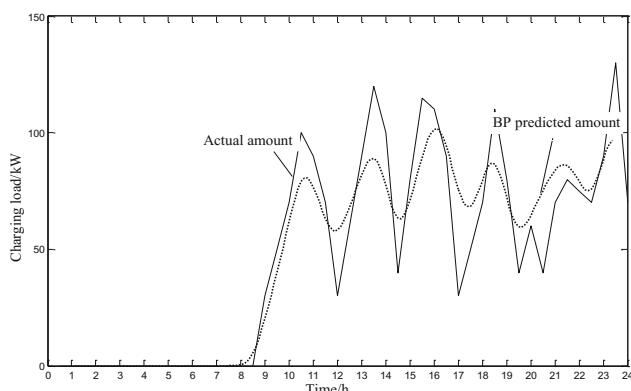
This article uses two indicators to decide the performance of the curve-fitting calculation by BP neural network and WNN, which are the relative error percentage and the prediction accuracy.



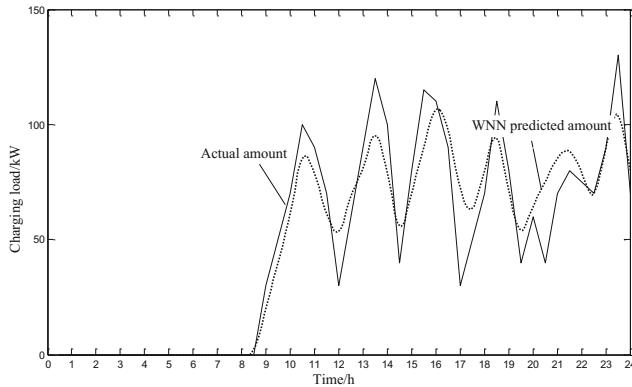
**Fig. 4.** Comparison of actual value and predicted value using BP neural network in June 16



**Fig. 5.** Comparison of actual value and predicted value using WNN in June 16



**Fig. 6.** Comparison of actual value and predicted value using BP neural network in June 21



**Fig. 7.** Comparison of actual value and predicted value using WNN in June 21

The relative error percentage is specified as follows:

$$\varepsilon = \frac{1}{T} \sum_{t=1}^T \left| \frac{P_t - P'_t}{P_t} \right| \times 100\% \quad (7)$$

Where  $\varepsilon$  is the relative error percentage;  $P_t$  is the actual amount of time  $t$ ;  $P'_t$  is the predicted amount of time  $t$ ;  $T$  is the sum of the entire amount, this place  $T = 96$ .

The prediction accuracy is specified as follows:

$$\varepsilon = \frac{1}{T} \sum_{t=1}^T \left| \frac{P_t - P'_t}{P_t} \right| \times 100\% \quad (8)$$

Where  $A$  is the relative error percentage;  $P_t$  is the actual amount of time  $t$ ;  $P'_t$  is the predicted amount of time  $t$ ;  $T$  is the sum of the entire amount, this place  $T = 96$ .

Then we could get the results as follows:

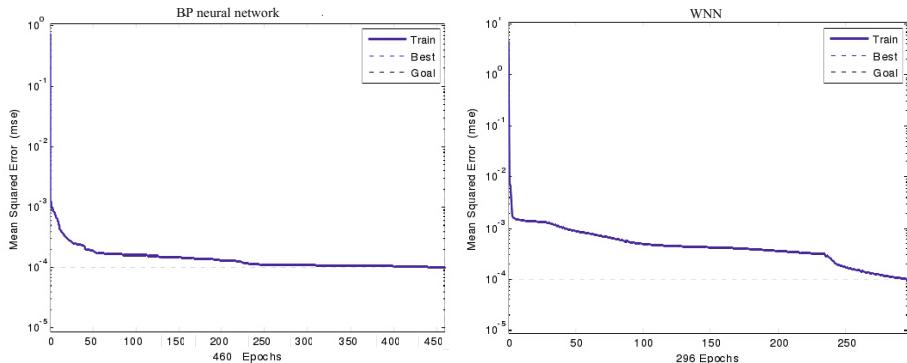
To the data in June 16, the relative error percentage and the prediction accuracy by BP neural network are 7.31% and 89.87%, while they are 6.50% and 92.34% by WNN.

To the data in June 21, the relative error percentage and the prediction accuracy by BP neural network are 8.60% and 88.17%, while they are 7.65% and 91.45% by WNN.

So we can see that the WNN is better than BP neural network in short-term load forecast of charging station from the above data.

And the training results of BP neural network and WNN of June 16 is as Fig 8. The training functions are sigmoid function and morlet function respectively.

From Fig 8 we can see that the training speed of WNN is faster than BP, so the computational complexity in the training of WNN is small, this ensures the model can cope with large amounts of data.

**Fig. 8.** Training results comparison of BP and WNN

To further explicate our conclusion the other two records are randomly selected to be the simulations, which are in December 16, 2013 (Monday) and June 21, 2013 (Saturday). And the Tab 1 gives a concise description of difference between the results from the WNN and BP neural network.

**Table 1.** Comparison of different forecasting methods

Simulations	relative error percentage		prediction accuracy		best training epochs	
	BP	WNN	BP	WNN	BP	WNN
June 16	7.31%	6.50%	89.87%	92.34%	460	296
June 21	8.60%	7.65%	88.17%	91.45%	526	299
December 16	7.53%	5.44%	89.15%	94.58%	504	275
December 21	7.63%	5.87%	88.58%	93.62%	482	281
Mean	7.77%	6.37%	88.94%	93.00%	493	288

We could get the conclusion that the relative error percentage of WNN is 1.40% lower than BP neural network and the prediction accuracy is 4.06% higher by comparing with four groups of data. And the best training epochs of WNN are very lower than BP neural network. So it's better to use WNN to get the short-term load forecast of charging station than BP neural network.

Based on the comparison with the forecast result from BP neural network, it is demonstrated that the convergence speed of the wavelet neural network is faster with more accurate precision.

Because of a few charging buses at this charging station and charging load is influenced by random factors seriously, it have a great fluctuation in load and the results differences of two methods aren't too obviously large. But it still proves that WNN is better than BP neural network in short-term load forecast of charging station.

## 5 Conclusion

On the basis of previous work, this paper uses the wavelet neural network (WNN) method to forecast the power load of charging station, analyzes the model characteristic and the performance of the WNN is compared with that of the conventional BP neural network. The simulations reveal that the main indicators relative error percentage and prediction accuracy of WNN are better than that of BP neural network in short-term load forecast of charging station. And the best training epochs of WNN are much smaller than BP neural network. It can provide scientific basis for the large-scale electric bus charging stations distributing rationally and orderly power load.

**Acknowledgments.** The authors are grateful to the support of the National High Technology Research and Development Program ("863" Program) of China (No. 2011AA05A114).

## References

1. Liu, K.: Comparison of very short-term load forecast techniques. *IEEE Trans. on Power Systems* 11(2), 877–882 (1996)
2. Zhang, D.: Study of power system load forecast based on artificial neural network and wavelet theory. Tianjin University, Tianjin (2002)
3. Xu, Z., Hu, Z., Song, Y., et al.: Coordinated charging of plug-in electric vehicles in charging stations. *Automation of Electric Power Systems* 36(11), 38–43 (2012)
4. Zhan, K., Song, Y., Hu, Z., et al.: Coordination of electric vehicle charging to minimize active power losses. *Proceedings of the CSEE* 32(31), 11–17 (2012)
5. Ge, W., Huang, M., Zhang, W.: Economic operation analysis of the electric vehicle charging station. *Transactions of China Electro Technical Society* 28(2), 15–21 (2013)
6. Zhang, D., Jiang, S., Bi, Y., Zou, G.: Study of power system load forecast based on wavelet neural networks. *Electric Power Automation Equipment* 23(8), 29–32 (2003)
7. Zhang, P., Pan, X., Xue, W.: Short-term load forecast based on wavelet decomposition, fuzzy gray correlation clustering and BP neural network. *Electric Power Automation Equipment* 32(11), 121–125 (2012)
8. Ge, S., Jia, O., Liu, H.: A gray neural network model improved by genetic algorithm for short-term load forecast in price-sensitive environment. *Power System Technology* 36(1), 224–229 (2012)
9. Yang, H., Wang, C., Zhu, K., et al.: Short-term load forecast based on phase space reconstruction and Chebyshev orthogonal basis neural network. *Power System Protection and Control* 40(24), 95–99 (2012)
10. Ying, C., Peter, B.L., Che, G., et al.: Short-term load forecast: similar day-based wavelet neural networks. *IEEE Trans. on Power Systems* 25(1), 322–330 (2010)
11. Che, G., Peter, B.L., Laurent, D.M., et al.: Very short-term load forecast: wavelet neural networks with data pre-filtering. *IEEE Trans. on Power Systems* 28(1), 30–41 (2013)
12. Wang, J., Wu, G., Li, Y.: Application of ant colony gray neural network combined forecast model in load forecast. *Power System Protection and Control* 37(2), 48–52 (2009)

13. Li, G., Zou, D., Tan, S.: Short-term load forecast for small power net based on chaos-artificial neural network theory. *Electric Power Automation Equipment* 26(2), 50–52 (2006)
14. Kenji, N.F., Anna, D.P., Carlos, R.M.: Short-term multimodal load forecast using a modified general regression neural network. *IEEE Trans. on Power Delivery* 26(4), 2862–2869 (2011)
15. Zhang, W., Xie, F., Huang, M., et al.: Research on short-term load forecast methods of electric buses charging station. *Power System Protection and Control* 41(4), 61–66 (2013)

# The Layout Optimization of Charging Stations for Electric Vehicles Based on the Chaos Particle Swarm Algorithm

Zhenghui Zhang, Qingxiu Huang, Chun Huang, Xiuguang Yuan,  
and Dewei Zhang

Electrical and Information Engineering, Hunan University, Changsha 410082, China

**Abstract.** Electric vehicle is an important part of the smart grid, and the location selection and the constant volume of charging stations for electric vehicles have been the research hotspot in the field of electric vehicles. In order to reasonably determine the scale and layout of charging station for electric vehicles, a novel model of the location selection and the constant volume of charging stations for electric vehicles considering time-space distribution ,power losses and the cost of new lines is established by taking the investment cycle costs and user convenience as indexes. Under the constraint of the related conditions, the objective function is constituted of the initial investment by the new station, network loss costs, the new line costs and electricity costs, and the target is to minimize the investment and user costs .Then the layout of charging station is optimized by the improved chaotic particle swarm algorithm; then chaotic sequence was formed and the corresponding relationship of variable range was optimized through logical mapping function. Example analysis shows that the proposed method has better convergence properties than the particle swarm optimization (PSO) algorithm, which can offer a new way for the layout of the electric vehicle charging stations.

**Keywords:** electric vehicle, charging station, Chaos Particle Swarm Optimization, size and layout, logical mapping function.

## 1 Introduction

To reduce fog and haze, reducing PM2.5 concentration, the development of energy-saving and environmentally friendly transport increasingly become the focus of public attention [1]. In this context, the electric car industry has been rapid development, and sales market has begun to take shape. Battery replacement can guarantee energy supplies quickly without compromising battery and help to extend battery life[2] That State Grid Corporation and China Southern Power Grid Company recommend changing the battery-based electric vehicle operations plan and develop optimal planning of electric vehicle battery charging station is imminent.

In [3], according to geographical factors and charging to determine the candidate station site station service radius, while regarding the investment, operating costs and

the cost of the charging station network losses minimum as the goal, the optimal mathematical model constructed electric vehicle charging station planning and adopt improved solving primal-dual interior point. In [4], services radius and traffic should be regarded as the main indicator of demand for charging stations, while it should also consider transportation, environmental protection and regional distribution capabilities and other external environmental conditions and the region's road network planning and construction planning. In [5], the author introduced the electric car stratification, partition scheduling philosophy, through the optimization of each agent in each period electric car charging and discharging the output load scheduling, allowing the system within the target time interval of the total variance of the minimum load level, the establishment of a network can be optimized based on double electric vehicle charging and discharging scheduling model. In [6], for electric vehicle charging load characteristics were analyzed in a simplified lithium battery I - U charging model, based on a single model to get the charging of electric vehicles, electric vehicle charging stations proposed two stage cluster model based on Poisson distribution and modeling day charging station load curve for electric car charging stations charging load agglomeration model.

This article drew the network loss and power line construction cost battery into charging station investment model, and established vehicle battery charging station location and size of new models, considering time and spatial distribution of electric vehicles, power loss and the cost of the new line. The model simulate electric vehicle traffic peak, the maximum extent possible to user need. Under the constraints of transmission power, reactive power compensation limit, the number of batteries and the station service radius, the initial investment by the new station, the net loss costs, new line costs and driving power costs constitute the objective function. For the model, using the chaos particle swarm optimization algorithm optimizes the layout of charging station.

## 2 Economic Model of the Battery Charging Station

The planning of battery charging station depends on its investment planning cycle cost and user convenience. Investment cycle costs are considered the whole process from the beginning of the construction of battery charging stations to battery charging stations putting the cost of the normal use User convenience aims at the user's shortest waiting time target in the battery charging stations and the shortest distance to battery charging stations. The shortest waiting time means that the battery charging station number, location and size of charging stations can satisfy the peak demand time, reducing or eliminating the queue probability of the user. Excluding the impact of traffic congestion and road maintenance and other uncertainties, with time-consuming and driving distance correspondence, the longer the distance traveled, the road takes longer, and vice versa. Thus, the road driving distance can be converted into power costs as a measure of user convenience indicators. That the study minimizes investment costs of battery charging station and power consumption costs with

the target determines the size of the battery charging station, location, and user services for power consuming. This is mathematical model:

$$\min G = \sum_{j=1}^n (G_{bj} + G_{cj}) \quad (1)$$

$$G_{bj} = \sum_{i=1}^4 G_{bij} \quad (2)$$

Where  $G$  is the annual cost of the battery charging station and  $G_{bj}$  is the investment costs of battery charging station  $j$ ;  $G_{cj}$  represents the power consumption costs of users who coming to the battery charging station  $j$ ,  $j = 1, 2, 3, \dots, n$  respectively representing the battery charging station number;  $G_{bij}$  is the  $i$  type of investment cost of the battery charging station  $j$ ,  $i$  values of 1, 2, 3, respectively representing the initial investment cost of battery charging stations, network loss costs, and the cost of the new line.

## 2.1 Investment Costs

### 2.1.1 The Initial Investment Cost

The initial investment cost of the new station  $j$  includes the purchase of equipment  $E_j$  and the cost of land acquisition costs  $A_j$ .  $E_j$  including the costs of AC charging, binning charger, battery replacement systems, power distribution monitoring charging, billing systems and emergency power monitoring charging machines and other equipment purchase , the number and capacity of the equipment associated with the size of the battery charging station.  $A_j$  is determined by the location and the area of battery charging stations. $G_{b1j}$  is calculated as follows:

$$G_{b1j} = \frac{r_0(1+r_0)^m}{(1+r_0)^m - 1} (A_j + E_j) \quad (3)$$

Where  $r_0$  is the investment recovery and  $m$  is the operating life of the battery charging stations.

### 2.1.2 Net Loss Expenses

Net loss costs  $G_{b2j}$  refer to the new battery charging station  $j$  access to the original grid, the additional costs caused by network outages. The cost associating with location of battery charging stations connected to the grid on is one indicator of charging station location's measuring net economic impacts.  $G_{b2j}$  is calculated as follows:

$$G_{b2j} = r_L \Delta A_j \quad (4)$$

Where  $r_L$  is the conversion coefficient of annual value of the net loss cost ,and  $\Delta A_j$  is the amount of incremental loss when new battery charging station  $j$  access to the original grid.

### 2.1.3 New Line Charges

New line costs  $G_{b2j}$  refers to the investment from new battery charging station to the nearest substation, and the fee is not only related with the location of the battery charging station but also the power load levels of regional planning.  $G_{b3j}$  is calculated as follows:

$$G_{b3j} = \frac{r_0(1+r_0)^m}{(1+r_0)^m - 1} \lambda_l L_j \quad (5)$$

Where  $\lambda_l$  is the investment cost of unit length of the double lines and  $L_j$  is the line length from battery charging station  $j$  to the nearest substation.

### 2.2 Driving Power Costs

The driving power cost is calculated as:

$$G_{cj} = \alpha \beta k w \sum_{i \in I_j} D_{ij} \quad (6)$$

Where  $\alpha$  is the tortuous coefficients of path,  $\beta$  is the roads open coefficient,  $k$  is the average number of charges per vehicle,  $D_{ij}$  is the distance from electric vehicle  $i$  to battery charging station  $j$ ,  $I_j$  is the collection of electric cars in the battery charging station  $j$ .

### 2.3 Constraints

To avoid affecting the security of the existing power grid, power quality and meet customer demand, the programming model of battery charging station need to add the following constraints:

1. The line transmission power of battery charging station is constrained.

$$p_j(t) \leq p_{j\max}(t) \quad (7)$$

Where  $p_j(t)$  stands for the transmission power between the battery charging station  $j$  and Transformer substation during the period ,while  $p_{j\max}(t)$  stands for the maximum value of  $p_j(t)$ .

2. The Boundary conditions of the reactive power compensation:

$$Q_{j\min} \leq Q_j \leq Q_{j\max} \quad (8)$$

Where  $Q_j$  stands for the compensation power of battery charging station  $j$  ,while  $Q_{j\max}$  and  $Q_{j\min}$  stands for the maximum and minimum value of  $j$ .

The constraints of battery quantity constraints:

$$n_j T_j \leq N_j \quad (9)$$

Where  $n_j$  represents the unit-hour service number of battery replacement system of  $j$ ,  $T_j$  represents the daily average working time of  $j$  and  $N_j$  represents the number of batteries of  $j$ .

Service radius constraints:

$$r_j \leq e_{soc} L_N \quad (10)$$

Where  $r_j$  represents the service radius of  $j$ ,  $e_{soc}$  represents the average charge state of the electric vehicle, and  $L_N$  represents each rated mileage of electric vehicle.

### 3 Chaos Particle Swarm Optimization

PSO originated in the simulated birds foraging behavior is an algorithm of matrix stochastic search by iteration [5]. The formula that particles update their velocity and position is presented as follow:

$$v_{i,j}^{k+1} = w v_{id}^k + c_1 p_{rand1}^k (x_{p,j} - x_{i,j}^k) + c_2 p_{rand2}^k (x_{g,j} - x_{i,j}^k) \quad (11)$$

$$x_{i,j}^{k+1} = x_{i,j}^k + v_{i,j}^{k+1} \quad (12)$$

Where  $c1$  and  $c2$  are the Learning factors,  $w$  is the inertia weight,  $x_{i,j}^k$  is the position of the particles in the cycle  $k$ ,  $x_{p,j}$  is the global optimal coordinates of the particles, and  $x_{g,j}$  is the locally optimal coordinates of the particles.PSO uses the speed and position simultaneously search model, having strong global optimization speed and capacity. However, PSO has high accuracy results easily and falls into local solution. To overcome these shortcomings, the chaotic motion having ergodicity, randomness and other characteristics is introduced PSO, when the particles fall into premature convergence, with chaotic disturbance escaping from local optima, and quickly find the optimal solution to improve the accuracy and convergence rate. Using chaotic sequence based on particle swarm optimization algorithm to obtain the desired results.

Mathematical process of chaotic motion is as follows: A random initial vector for an  $N$ -dimensional vector

$$x_0 = [x_{0,1}, x_{0,2}, \dots, x_{0,N}]^T, x_{0,n} \in [0,1]$$

According to the model of chaotic sequence, begin iteration. According to reference [10], we use the logic mapping function to obtain the iterative sequence  $x_{m,n}$ . The mathematical expression is calculated as follows:

$$\begin{aligned} x_{i+1,j} &= 1 - 2x_{i,j}^2 \\ i &= 0, 1, 2, \dots, M; j = 0, 1, 2, \dots, N \end{aligned} \quad (13)$$

Here are the steps of chaotic motion;

According to formula (13), the model generates a  $N \times 2$  matrix originated from the random initialization in the region  $[-1, 1]$ .

The variables of chaotic motion are mapped into population particle according to the following formula:

$$y_{i,j} = x_{g,j} + R_{i,j} (2x_{i,j} - 1) \quad (14)$$

Where chaotic variables  $x_{i,j}$  are transformed into the round which regards the current global optimum location  $x_{g,j}$  as the center and  $R_{i,j}$  as the radius.  $R_{i,j}$  is the radius of the chaotic motion of each neighborhood, so the range of particle populations  $y_{i,j}$  optimized by the logic self-mapping functions as follows:

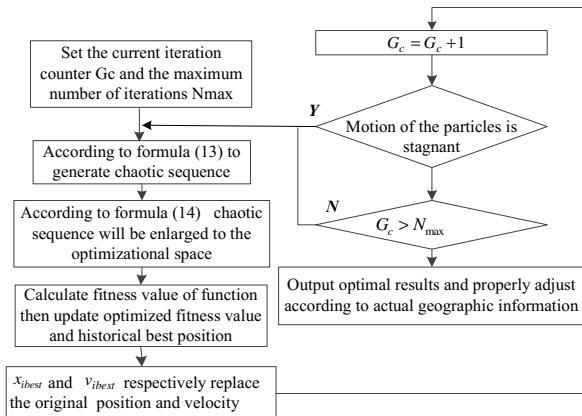
$$y_{i,j} \in [x_{i,j} - R_{i,j}, x_{i,j} + R_{i,j}] \quad (15)$$

Calculate the fitness value  $f(x_{m,n})$  of the objective function ,update the global optimum value and global best position ,and update the chaotic iteration historical optimal fitness value  $f(g_{best})$  and best position  $x_{ibest}$ . If the global iteration optimal fitness value after chaotic iteration is better than the global historical optimum fitness value  $G_{best}$ , then replace the original best position and velocity with the position and velocity of the chaotic iteration.

Wherein the speed is calculated as:

$$v_{ibest} = \frac{x_{ibest} - x_i}{\|x_{ibest} - x_i\|} \quad (16)$$

In summary, the planning flow chart of battery charging station based on CPSO planning flow chart below:



**Fig. 1.** Planning process of charging station based on CPSO

It should be noted that:

$N_i$  represents the number of charging stations, when the construction scale meet user needs of the planning area;  $N_{max}$  represents the maximum number of charging stations demand.  $N_i$  is calculated as follows:

$$N_i = \frac{C_{max}}{\alpha_1 \alpha_2 \beta \cos \varphi l_i} \quad (17)$$

Where  $C_{max}$  represents the number of the electric vehicle needs within a maximum charging time,  $\alpha_1$  represents the ratio of the charging station while working,  $\alpha_2$

represents meanwhile rate of charger,  $\beta$  represents the charging efficiency of charger, and  $\cos\phi$  indicates the power factor; According to "State Grid Corporation of electric vehicle charging facilities guidance", the service ability of electric vehicle charging stations are divided into different levels,  $i = 1, 2, 3$ ;  $L_i$  indicates the level of service capabilities  $i$  of the charging stations.

Determine the standard whether the particle motion is stagnant:

$$\Delta G_i = \frac{G_i - G_{pbest}}{G_i} < \delta \quad (18)$$

$$N_{\Delta G} < N_c \quad (19)$$

Where  $\delta$  and  $N_c$  are the both constant set according to the actual situation, the initial value of  $x$  is set to 0 .If  $\Delta G_i \geq \sigma N_{\Delta G} = N_{\Delta G} + 1$ .If it do not satisfy the formula (18), the particle motion is in stagnation, and re-enter the chaotic motion.

## 4 Cases Considered

A district with a total area of 70.3 km<sup>2</sup>, the resident population of people is 10.7 million , stuff span is 11.06 km, north-south span is 9.12 km; The district has two main roads, four secondary roads, 20 branches, and is divided into four main roads partitions; The total daily traffic is 19,000.The plot is well-developed industrial and tourism, electric buses and taxis account for a larger proportion of the total electric cars, and specific basic data shown in Table 1.The plot focused on residential, tourist and commercial areas, such as various types of land premium detailed configuration Table 2, different sizes of the charging station equipment as shown in Table 3.As the reality of electric vehicles type, service time-consuming, tortuous road factor, flow coefficient is not the only way to simplify the problem, taking the average of the calculated numerical example, where the value of the parameters as follows:  $r_0 = 0.1$ ;  $r_1 = 0.2$ ;  $\lambda_1 = 16500 \text{ yuan/km}$  ;  $k = 242$ ;  $\alpha = 1.05 \sim 1.45$  ;  $\beta = 1.0 \sim 1.3$  ;  $m = 20$  ;  $w = 1.1 \text{ yuan/km}$  ;  $T = 10h/a$  ;  $e_{soc} = 50\% \sim 75\%$  ;  $L_N = 150 \sim 200 \text{ km}$  .

**Table 1.** Basic data

Comparison Project	Number/d	Rechargeable Battery( set/d)	The maximum time demand for electricity
Bus	500	2.0	10:00~11:00,20:00~21:00
Taxi	3500	3.0	12:30~13:30,21:00~22:00
Social vehicle	16000	0.1	18:00~19:00

**Table 2.** Land price yuan/m<sup>2</sup>

Industrial land	Residential land	Commercial land	Service land
800	3200	5 000	900

**Table 3.** Detailed configurations of charging station with different scales

Scale	Service capabilities (No./d)	Battery Replacement System/set	Charging pile/	Charger/	Covers/ m <sup>2</sup>
1	550	6	100	85	14 000
2	550	4	80	68	9 200
3	300	2	50	50	4 500

According to Table 1 for electricity demand of the battery pack, charging stations must be built to ensure that 5400 set of batteries can service every day. In conjunction with Table 2 and equation (17), when the battery charging station at the scale of the first configuration requires a minimum of six, When the battery charging station by the third configuration, up to 18 battery charging stations. When n (6~18)is the loop variable to calculate the total cost of the current number of battery charging stations, the minimum value corresponding to the optimal planning results as the result of the size and layout of choice.

Ideally, the initial device configuration is the main factor causing different investment costs of the battery charging station of different sizes; increasing as the number of the battery charging station while its size decreases, a single battery replacement system increases the proportion accounted for equipment investment. The battery replacement system's influence on the initial construction costs is the largest, therefore, the more battery charging stations, the higher and the total investment. That example of the base 100 run independently calculates the minimum cost of construction 6-18 Block battery charging station as follows: 30.702, 30.764, 30.627, 30.559, 30.631, 30.705, 30.784, 30.846, 30.971, 31.100, 31.223, 31.351 million yuan. Therefore, under the given conditions, when the planning area build 9 battery charging stations, the results are optimal .Each optimized location and size of charging station is in Table 4, The investment costs of the battery charging station, power costs and time are shown in Table 5, wherein the time of power consumption on the road is calculated on the basis of the assumption that the speed of vehicle traveling is maintained at 100km/h .The results show that the results optimized from the algorithm not only guarantee a minimum investment costs, but also meet the user's convenience. Figure 2 shows the dynamic evolution of optimal planning when using CPSO algorithm and PSO algorithm.

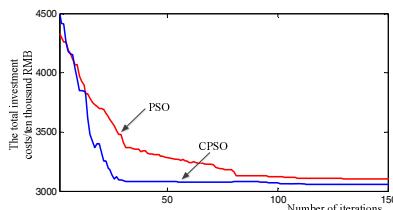
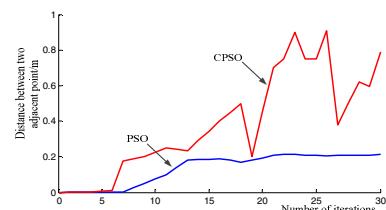
**Table 4.** Optimization results of charging station location

Number	Coordinate		Scales
	x/km	x/km	
1	1.422	2.060	3
2	1.264	2.270	2
3	1.350	2.733	2
4	1.710	1.513	3
5	0.645	2.046	3
6	0.253	0.294	2
7	1.521	1.296	3
8	1.990	1.603	2
9	1.253	2.681	1

**Table 5.** Simulation results of charging station optimization(Ten thousand yuan)

Number	Road time-consuming/min	The initial investment in equipment	Land costs	Net loss expenses	New line charges	Power costs
1	16.0	140.0	45.1	40.1	75.1	9.92
2	10.8	45.1	82.2	90.0	141.3	11.57
3	13.1	44.5	82.2	90.0	141.3	11.57
4	12.2	82.2	45.1	40.1	75.0	9.92
5	9.8	82.2	45.1	40.1	75.0	9.92
6	12.2	45.1	82.2	90.0	141.0	11.57
7	11.0	82.0	45.1	40.2	75.0	9.92
8	12.4	82.0	82.0	89.9	141.2	11.57
9	10.3	45.1	140.0	137.9	251.3	14.86

Figure 3 shows the changes in distance between two adjacent tracks when using CPSO algorithm and PSO algorithm .Figures 2 and 3 can be seen: CPSO algorithms avoid a premature, convergence speed and optimal results are better than other three PSO algorithms. The results showed that the size of the battery charging stations election and layout of the model established in this paper while reducing construction costs and investment; On the other hand can meet different traffic densities for electricity demand, in line with expectations envisaged.

**Fig. 2.** Optimal planning of dynamic evolution process of CPSO and PSO**Fig. 3.** Change trajectory between two ads points in CPSO and PSO

## 5 Conclusion

This paper presents a mathematical model which not only considering the cost of the investment, but also adding user convenience. The model meets the needs of the construction side and consumers, and reflects the essence of the charging station programming problem from the comprehensiveness and economy.

CPSO algorithm used sensitivity to initial value and ergodicity of chaos to initialize the population; then by logical self-mapping function, the chaotic sequence is formed and the corresponding relationship of variable range is optimized. Initialized by chaotic PSO algorithm can make a good initial value from optimization, while stagnant standard iterative update process to make the search more precise, the PSO algorithm to overcome the premature, into local extreme defects. CPSO algorithm can

start optimization from a good initial value through Chaos initialization, while stagnant standard of iterative update process can make the search more precise, overcome the prematurity of the PSO algorithm and defects that PSO fall easily into local extreme.

As electric vehicle battery charging station was planned in a district, optimization model presented in this paper is proved to be scientific and feasible.

**Acknowledgments.** The authors are grateful to the support of the National High Technology Research and Development Program ("863" Program) of China (No. 2011AA05A114).

## References

1. Xu, J., Ding, G., Yan, P., et al.: Compositional characteristics in Beijing PM (2.5) and source analysis. *Journal of Applied Meteorology* 18(5) (2007)
2. Gao, C., Zhang, L., Xue, F., et al.: Grid planning considering capacity and site of large-scale centralized charging stations. *Proceedings of the CSEE* 32(7), 40–46 (2012)
3. Ren, Y., Shi, L., Zhang, Q., et al.: Optimal Distribution and scale of charging stations for electric vehicles. *Automation of Electric Power Systems* 35(14), 53–58 (2011)
4. Xu, F., Yu, G., Gu, L., et al.: Tentative analysis of layout of electrical vehicle charging stations. *East China Electric Power* 37(10), 1678–1682 (2009)
5. Yao, W., Zhao, J., Wen, F., et al.: A charging and discharging dispatching strategy for electric vehicles based on bilevel optimization. *Automation of Electric Power Systems* 36(11), 30–37 (2012)
6. Liu, Z.P., Wen, F.S., Ledwich, G.: Optimal planning of electric-vehicle charging stations in distribution systems. *IEEE Transactions on Power Delivery* 28(1), 102–110 (2013)
7. He, J., Zhou, B., Feng, C., et al.: Electric vehicle charging station planning based on multiple-population hybrid genetic algorithm. In: 2012 International Conference on Control Engineering and Communication Technology, pp. 403–406. IEEE, Liaoning University (2012)
8. Hu, Z., Song, Y., Xu, Z., et al.: Impacts and utilization of electric vehicles integration into power systems. *Proceedings of the CSEE* 32(4), 1–10 (2012)
9. Ji, Z., Liao, H., Wu, Q., et al.: Particle swarm optimization and application, pp. 28–45. Science Press, BeiJing (2009)
10. Liu, C., Ye, C.: Mutative scale chaos particle swarm based on the mapping logic function. *Application Research of Computers* 28(8), 25–27 (2011)
11. Tian, W., He, J., Jiang, J., et al.: Multi-objective optimization of charging dispatching for electric vehicle battery swapping station based on adaptive mutation particle swarm optimization. *Power System Technology* 36(11), 25–29 (2012) (in Chinese)

# An Improved Feature Weighted Fuzzy Clustering Algorithm with Its Application in Short-Term Prediction of Wind Power<sup>\*</sup>

Xinkun Wang<sup>\*\*</sup>, Diansheng Luo, and Hongying He

College of Electrical Information and Engineering, Hunan University,  
Changsha 410082, Hunan, China  
S12092099@hnu.edu.cn

**Abstract.** Based on improved feature weighted fuzzy clustering and Elman neural network, short-term forecasting method of wind power is proposed in the paper. Because physical properties of wind identify wind types with different importance, the paper introduces weighted factor in traditional FCM fuzzy clustering algorithm and synthetically clusters the data samples of historical wind type. Aim at clustering results, it dynamically establishes model of Elman neural network in order to predict wind power output value of the same clustering results in target day. Furthermore, the paper simulates experiments with measured data of a domestic wind field, which proves the superiority and practicability of the proposed method.

**Keywords:** Wind power prediction, wind type, feature weighted fuzzy clustering, Elman neural network.

## 1 Introduction

Wind energy is renewable clean energy with high commercial value. Researched and applications of wind power has attracted the world's attention. But compared with characteristics of other conventional energy, wind power with clearance is random, and not completely controllable [1-2]. As more and more large-scale wind field grid power generation, the proportion of wind power in power grid is rising year by year. Therefore, if it can accurately forecast power of wind power, it will greatly improve the safety and stability of power system operation [3], provide powerful guarantee for electric power dispatching departments to make reasonable scheduling plan, and minimize cost of power system operation [4-5].

Existing researches have shown that the factors that affect wind speed and wind power prediction accuracy mainly includes physical properties of forecast object, historical data sample pretreatment technology, prediction model, etc [6]. In order to

---

<sup>\*</sup> Fund project: the national natural science foundation of China(51277057); Research of load forecasting theory and method under smart grid environment.

<sup>\*\*</sup> Corresponding author.

effectively extract historical data samples with better similarity, Zhiyong Ding [7] applied traditional Euclidean distance as measure, took wind speed as only index of clustering direction. By two times of clustering, they divided the whole year into several types of similar periods in a row. Yangyang Meng [8] used the correlation coefficient to calculate similarity among data samples. When calculating the sample similarity, they comprehensively considered various physical quantities of wind. However, if they could distinguish the influence degrees of each physical property on sample similarity computing, it would get better data sample clustering results.

This paper puts forward a data preprocessing technology based on feature weighted fuzzy clustering. Considering the differences of various physical attribute indexes of wind, it introduces feature weighting factor to represent the physical characteristics index of the wind to judge the influence degree of the comparability of data sample. In addition, the paper integrates the history data sampling with fuzzy clustering. After completing fuzzy clustering of data sample, the paper employs improved dynamic Elman neural network prediction model. It uses trained prediction model to predict wind power output of target days in clustering results. The paper verifies experimental data of a wind farm in 2012, proving that the prediction accuracy of proposed method is improved greatly and the method has strong practicability.

## 2 Data Preprocessing Technology Based on the Fuzzy Clustering

Clustering analysis is a kind of multi-element statistic analysis, also is the supervision and an important branch of pattern recognition [9]. Implementation method can be roughly divided into four types: hierarchical clustering method, clustering method based on equivalent relation, graph clustering method and fuzzy clustering method based on objective function [10-11]. The former three methods are difficult to deal with abundant data, without strong practicality. The fuzzy clustering method based on the objective function can simplify the data sample cluster analysis as optimal solution problem with nonlinear constrained programming [12], which is easier to implement by computer receiving wide applications.

### 2.1 The Structure and Normalized Processing of Data Samples Clustering

In the study of wind power prediction, when choosing physical properties indexes of wind data sample of the history days, it needs to consider wind type related indicators, including wind speed, wind direction, temperature, pressure, humidity and the surface sensible heat, etc. If looking all these indicators as input variables neural network, it will lead to reduce efficiency of neural network calculation [13]. Hongtao Shi[14] pointed out that the correlation of pressure, humidity, and surface sensible heat are much less related with wind power in the wind power output prediction. Accordingly, we can ignore the influence of these three indicators. To reduce the computing time of prediction model, the paper chooses the appropriate number of input variables, without affecting the accuracy of prediction. Hence, it selects wind speed, wind direction

and temperature to construct the historical wind data samples and fuzzily cluster the samples. Sample structure form is as follows.

$$x_i = [x_{i1}(m), x_{i2}(m), x_{i3}(m)] \quad (1)$$

where  $m=1,2,3,\dots,n$ ,  $x_{i1}(m)$  is wind speed value at moment  $m$  in the  $i$ th day,  $x_{i2}(m)$  is wind absolute value at moment  $m$  in the  $i$ th day,  $x_{i3}(m)$  is temperature value at moment  $m$  in the  $i$ th day.

Due to the dimension difference of each physical quantity, it will affect the final classification result if directly clustering data samples. On the other hand, normalization on the original data samples can effectively reduce the impact. This article adopts the method of range transformation, transforming the sample value between the interval  $[0, 1]$ , namely, the following conversion formula.

$$x_{ij}(m)' = \frac{x_{ij}(m) - \min(j)}{\max(j) - \min(j)} \quad (2)$$

where  $x_{ij}(m)$  is original sample values,  $\min(j)$  and  $\max(j)$  are minimum and maximum data samples of the  $j$ th index in the  $i$ th day, respectively.

## 2.2 Improved Feature Attribute Weighted Fuzzy Clustering Algorithm

In the traditional FCM clustering algorithm, in order to show the similarity between sample and the  $i$ th sample center  $v_c$ , usually, we calculate Euclidean distance of physical quantities samples values and make simple summation. Euclidean distance is  $d_{ci}^2(x_i, v_c) = \|x_i - v_c\|^2$ . This paper directly adopts FCM clustering algorithm, to some extent, it ignores the differences in physical quantities of wind speed, wind direction, temperature, etc. Moreover, these differences reflect that various physical quantities of wind are uneven on the clustering result of "contribution" degree [15]. Considering the differences of physical quantities in wind type have different impacts on clustering results, this paper introduces characteristics attribute weighted vector  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  to show "contribution rate" of various factors on clustering in the sample space.

Suppose there is a cluster sample set  $X = [x_1, x_2, \dots, x_N]$ , in which each sample  $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]$  contains  $n$  physical attributes. The number of initial clustering centers is  $C$ , namely, clustering centers sets are  $v_i = [v_1, v_2, \dots, v_c]$ .

Define improved new objective function as follows.

$$J(U, V) = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^2 \left\{ \sum_{p=1}^n \omega_p (x_{jp} - v_{ip})^2 \right\} \quad (3)$$

where  $u_{ij}$  is membership degree, which indicates the membership degree of data sample and clustering center.  $\omega_p$  is weighted vector value of each physical quantity in data samples.

The distance expression of improved data sample similarity is as Eq. (4).

$$d_{ij}^{*^2} = \sum_{p=1}^n \omega_p (x_{jp} - v_{ip})^2 \quad (4)$$

The objective function applies Lagrange multiplier method to solve the membership degree, which is shown in Eq. (5).

$$u_{ij}^* = \left\{ \sum_{k=1}^c \left( \frac{d_{ij}^{*^2}}{d_{kj}^{*^2}} \right)^{\frac{1}{m-1}} \right\}^{-1} \quad (5)$$

Update the cluster center as follows.

$$v_{ip} = \frac{\sum_{j=1}^N u_{ij}^m \cdot \omega_p \cdot x_{jp}}{\sum_{j=1}^N u_{ij}^m \cdot \omega_p} \quad (6)$$

In order to measure the effectiveness of the proposed fuzzy clustering, we define partition coefficient [16] in Eq. (7).

$$F(U, c, \omega) = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^N (u_{ij})^2 \quad (7)$$

If  $F = 1$ , hard division is established, and fuzzy clustering division is  $F < 1$ . Generally, we are always willing that fuzzy partition is as clearly as possible, namely, the value of  $F$  is the larger, the better.

Updating weighted vector  $\omega$  uses the following iterative formula:  $\omega_i^* = \omega_i + N(0, \gamma\delta)$ ,  $i = 1, 2, \dots, s$ , where  $\delta = 1 - F(U, c, \omega)$  is learning rate,  $N = (0, \sigma^2)$  is a Gauss random variable. The preceding analysis shows that when the clustering validity coefficients tend to be 1,  $\delta$  tends to 0, the Gauss random variable tends to be 0 as well, which means the value of weighted vector  $\omega$  tends to be stable or convergent.

In order to get favourable clustering results, we need to minimize objective function  $J(U, V)$ . Improved feature attribute weighted fuzzy clustering algorithm process is as follows.

Step 1: determine the initial clustering center  $V_i^{(L)}$ . Meanwhile, determine the number of initial clustering types C, data sample N,  $\epsilon$  and iteration times  $L = 0$ .

Step 2: set the initial weighted vector  $\omega_0$ , then update membership degree matrix through the initial clustering center and the formulas (4) and (5).

Step 3: according to the formula (6), update clustering center, get the  $V_i^{(L+1)}$ . Combining fuzzy clustering partition coefficient, update  $\omega$  according to the weighted vector iteration formula.

Step 4: if  $\|V_i^{(L+1)} - V_i^{(L)}\| < \epsilon$ , stop the algorithm, output membership degree matrix and cluster center at the same time. Otherwise, let  $L = L + 1$ , repeat step 1 to step 4.

### 3 The Wind Power Prediction Model Based on Elman Regression Neural Network

The input and output of neural network prediction model is a kind of highly nonlinear mapping relationship. We can close to any nonlinear function in an arbitrary precision by adjust the connection weights of neural network, input and output of network and the hidden layer node number. Therefore, it is able to solve problems of complex nonlinear, uncertain intellectual, uncertainty [17-18]. Currently, in the study of wind power short-term prediction, BP neural network is the most widely used [19]. However, the traditional BP neural network is a kind of static multilayer feed-forward neural network. Using Static feed-forward network to modeling and forecasting for dynamical system is in effect to change the dynamic time domain into static time domain problems. This will inevitably encounter many unknown problems. In the short-term forecast of wind power, it can objectively and directly reflecting the dynamic characteristics of the system if directly adopt the dynamic neural network modeling. Elman regression neural network is a kind of typical dynamic neural network [20]. It is based on the basic structure of the BP artificial neural network, by storing the internal state to make it has the function of each map dynamic characteristics, consequently, make the system have the ability to adapt to the time-varying characteristic [21].

#### 3.1 The Principle and Algorithm of Elman Neural Network

The nonlinear state space expression of Elman neural network is expressed as following.

$$\begin{cases} y(k) = g(w^3 x(k) + b_2) \\ x(k) = f(w^1 x_c(k) + w^2(u(k-1)) + b_1) \\ x_c(k) = x(k-1) \end{cases} \quad (8)$$

Here,  $k$  represents time,  $y$ ,  $x$ ,  $u$ ,  $x_c$ , denote one-dimensional vector output node, 1-dimensional unit vector of the hidden layer node, n-dimensional input vector and m-dimensional feedback state vector;  $w^3$ ,  $w^2$ ,  $w^1$  denote the connection weights matrix of matrix layer to output layer, input layer to hidden layer and associated layer to hidden layer;  $b_1$ ,  $b_2$  denote the value of the input layer and the hidden layer respectively;  $f(x)$  is the transmitted function of hidden layer neurons, and it always takes sigmoid function.  $g(x)$  is the transmitted function of output layer, it uses pure line function. Let the system's actual output of step  $k$  is  $y_d(k)$ , then the objective function: error function of Elman network could be expressed as Eq.(9).

$$E(k) = \frac{1}{2} (y_d(k) - y(k))^T (y_d(k) - y(k)) \quad (9)$$

According to the gradient descent method, calculate  $E(k)$  the partial derivative of right value and make it 0, and could obtain the learning algorithm of Elman neural network.

$$\begin{cases} \Delta w_{ij}^3 = \eta_3 \delta_i^0 x_j(k) \quad (i=1,2\cdots,m; j=1,2\cdots,n) \\ \Delta w_{iq}^2 = \eta_2 \delta_j^h u_q(k-1) \quad (j=1,2\cdots,n; q=1,2\cdots,r) \\ \Delta w_{jl}^1 = \eta_1 \sum_{i=1}^m (\delta_i^0 w_{ij}^3) \frac{\partial x_j(k)}{\partial w_{jl}^1} \quad (j=1,2\cdots,n; l=1,2\cdots,n) \end{cases} \quad (10)$$

$$\delta_i^0 = (y_{di}(k) - y_i(k)) g_i'(\cdot) \quad (11)$$

$$\delta_j^h = \sum_{i=1}^m (\delta_i^0 w_{ij}^3) f_i'(\cdot) \quad (12)$$

$$\frac{\partial x_j(k)}{\partial w_{jl}^1} = f_i'(\cdot) x_l(k-1) + \alpha \frac{\partial x_j(k-1)}{\partial w_{jl}^1} \quad (13)$$

$$(j=1,2,\dots,n; l=1,2,\dots,n)$$

where  $\eta_1$ ,  $\eta_2$ ,  $\eta_3$  are step length of learning of  $\omega^1$ ,  $\omega^2$ ,  $\omega^3$ , respectively.

## 4 Simulation and Results Analysis

This paper makes experiments by taking the measured data of a wind far in 2012 as study object. There are 20 fans in the wind far, and each fan has an installed capacity rating of 9500 MW. After clustering data samples, we select the data sample of December to November clustering results as a training sample in the Elman neural network model. As for a clustering result, we take the actual power output samples of past days as the outputs of the predictive model, and we select  $n$  similar days in the same clustering results, and take their actual power output samples as the inputs while training the Elman neural network model. The date of November 21st in this clustering result is selected as the target day. The selected clustering result includes 15 samples, and it samples once every 10 minutes, that is, each data samples 144 points every day. This paper conducts comparison of prediction about simple BP neural network methods, simple Elman neural network method and Elman neural network method based on feature weighted clustering method.

In order to verify the practicability of prediction method, we apply the root mean square error ( $E_{RMSE}$ ), average relative error ( $E_{MRE}$ ), maximum error ( $E_{MAX}$ ) to make quantitative analysis for each prediction method.

Root mean square error:

$$E_{RMSE} = \sqrt{\left[ \sum_{i=1}^n \left( \frac{x_i' - x_i}{x_i} \right)^2 \right] / n} \quad (14)$$

The average relative error:

$$E_{MRE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i' - x_i}{x_i} \right| \quad (15)$$

Maximum error:

$$E_{MAX} = \max(|x_i' - x_i|) \quad (16)$$

where  $n$  is the number of prediction sample,  $x_i'$  is prediction power output value,  $x_i$  is the measured power output value.

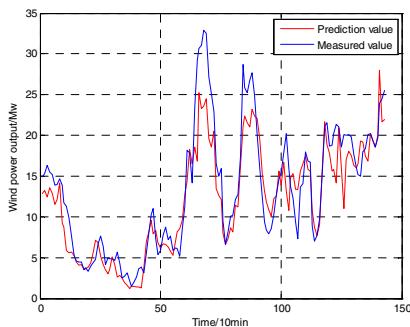


Fig. 1. Result in pure Elman neural network

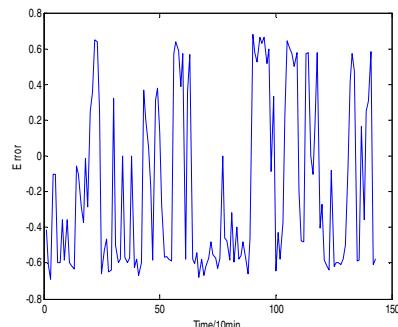


Fig. 2. Error in pure Elman neural network

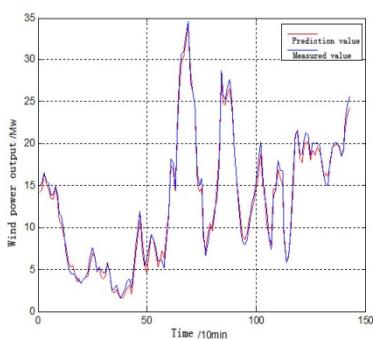


Fig. 3. Prediction results of proposed method

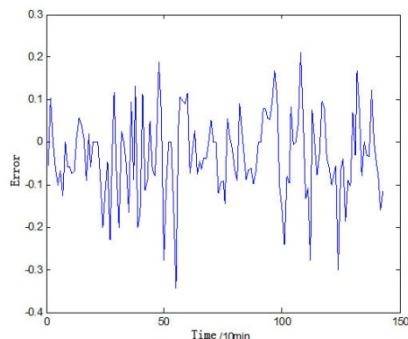


Fig. 4. Prediction error of proposed method

**Table 1.** Error statistics of different prediction methods

Methods	$E_{RMSE}/\%$	$E_{MRE}/\%$	$E_{MAX}/\%$
BP neural network	25.13	19.45	76.35
Elman neural network	22.72	14.63	61.42
Characteristics clustering Elman neural network	10.95	7.07	34.29

Fig. 1 ~ Fig. 4 compare the proposed method and the simple Elman neural network method in terms of their prediction result and prediction error. The training sample of simple Elman neural network method is the data sample in the first 10 consecutive days before December 16th. As shown in Fig. 1 and Fig. 2, the fluctuating range of the prediction error of simple Elman neural network method is higher, and the whole prediction accuracy is not ideal though its prediction results are similar with actual values. As for the proposed method, the number of the samples whose prediction error is less than 12% is 118, accounted for 81.94% of the total predicted data. And the number of those whose prediction error is more than 20% accounts for 6.29% of the total predicted data, which is just a small part. Therefore, we can cluster samples with high similarity degrees effectively and correctly after weighting feature, and thus improves prediction accuracy significantly.

Table 1 shows the statistical error of each prediction method. It's clear to see that the effect of the various prediction methods arranged in descending order is: simple BP neural network, simple Elman neural network, Elman neural network based on feature weighted clustering method. The maximum error in simple BP neural network is double of the proposed method. By modeling data samples which get after the dynamic Elman neural network is clustered, we can better match the dynamic characteristics of wind power generation system, so the proposed method can guarantee stability, high accuracy of predictions with high practicality.

## 5 Conclusion

This paper selects physical characteristics such as historical day wind speed, wind direction, temperature as data samples, and classifies the similar data samples as a class through feature weighted data sample pre-processing method based on fuzzy clustering. On the basis of same clustering results, the paper creates dynamic Elman neural network and makes short-term prediction of wind power. Conclusions are as follows: (1) The proposed clustering method can effectively classify highly similar data samples as a class and get better clustering effect. These highly similar training data samples can effectively improve stability and accuracy of prediction model. (2) By comparison with the traditional static neural networks, dynamic Elman neural network has better dynamic characteristics of wind power generation system. Compared with the conventional BP neural network in terms of prediction error, Elman neural network shows dynamic and practical advantages in wind power prediction. (3)

The paper uses measured data of a wind farm in 2012 to verify the proposed method.

The accuracy of prediction results of output power can reach 89.1%, which could meet National GridQ / GDW 588—2011 function specification of wind power forecasting requirements. Compared to other prediction methods, the proposed method has higher stability and accuracy to some extent.

## References

1. Lei, Y., Wangk, W., Yin, Y., et al.: Value analysis of wind power on power system operation. *Grid Technology* 26(5), 10–14 (2002)
2. Du, Y., Lu, J., Li, Q., et al.: Wind power short-term wind speed forecasting based on minimum squares support vector machine. *Grid Technology* 32(15), 62–66 (2008)
3. Fabbri, A., GomezSanRoman, T., RivierAbbad, J., et al.: Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Transactions on Power Systems* 20(3), 1440–1446 (2005)
4. Liu, Y., Han, S., Hu, Y.: Research overview of wind power farm output in the short term forecast. *Modern Power* 24(5), 6–11 (2007)
5. Yang, X., Xiao, Y., Chen, S.: Wind speed and generated power forecasting wind farm. *Proceedings of the CSEE* 25(11), 1–5 (2007)
6. Zhou, S., Mao, M., Su, J.: Wind power prediction based on principal component analysis and artificial neural network. *Grid Technology* 35(9), 128–132 (2011)
7. Ding, Z., Yang, P., Yang, X., et al.: Support vector machine wind power prediction method based on continuous time clustering. *Automation of Electric Power Systems* 36(14), 131–135 (2012)
8. Wind power short-term forecast based on similar days and artificial neural network. *Grid Technology* 34(12), 163–167 (2010)
9. Zadeh, L.A.: Fuzzy logic= computing with words. *IEEE Transactions on Fuzzy Systems* 4(2), 103–111 (1996); And artificial neural network based on similar days of wind power short-term forecast
10. Ruspini, E.H.: A new approach to clustering. *Information and Control* 15(1), 22–32 (1969)
11. Gao, X.: Analysis and application of fuzzy cluster. Electronic science and technology of Xi'an university press (2004)
12. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
13. He, D., Liu, R.: Dynamic integration of wind power short-term forecast of neural network based on principal component analysis. *Power System Protection and Control* 41(4), 50–54 (2013)
14. Shi, H., Yang, J., Ding, M., et al.: Short-term wind power prediction method based on the wavelet-BP neural network. *Automation of Electric Power Systems* 35(16), 44–48 (2011)
15. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems* 3(3), 370–379 (1995)
16. Bezdek, J.C.: Cluster validity with fuzzy sets (1973)
17. Thanasis, G.B., Theocharis, J.B.: Long-term wind speed and power forecasting using local recurrent neural network models. *IEEE Trans. on Energy Conversion* 21(1), 273–284 (2006)

18. Lee, K.Y., Cha, Y.T., Park, J.H.: Short-term load forecasting using an artificial neural network. *IEEE Transactions on Power Systems* 7(1), 124–132 (1992)
19. Alexiadis, M.C., Dokopoulos, P.S., Sahsamanoglou, H.S.: Wind speed and power forecasting based on spatial correlation models. *IEEE Transactions on Energy Conversion* 14(3), 836–842 (1999)
20. Zhu, M., Wen, C.: Short-term load forecasting of Elman neural network based on meteorological factors. *Electric Power System and Automation* 17(1), 23–26 (2005)
21. Aussem, A.: Dynamical recurrent neural networks towards prediction and modeling of dynamical systems. *Neurocomputing* 28(1), 207–232 (1999)

# Charging Load Forecasting for Electric Vehicles Based on Fuzzy Inference<sup>\*</sup>

Jingwei Yang, Diansheng Luo<sup>\*\*</sup>, Shuang Yang, and Shiyu Hu

The College of Electric and Information Engineering, Hunan University,  
Changsha 410082, China

**Abstract.** Large scale of electric vehicles (EVS) integration will pose great impacts on the power system, due to their disorderly charging. Electric cars' charging load cannot be forecasted as the traditional power load, which is usually forecasted based on historical data. There need to be some other methods to predict electric vehicles charging load, in order to improve the reliability and security of the grid. This paper analyze the travel characteristics of electric vehicles, then use the fuzzy inference system to emulate the process of drivers' decision to charge their cars, the charging probability is attained in the given location. Finally, the daily profile of charging load can be predicted according to the numbers of electric vehicles forecasted in Beijing.

**Keywords:** electric vehicle, travel characteristics, charging load, fuzzy inference.

## 1 Introduction

Automobile driving consumes large amounts of oil resources and emits large amount of fumes, makes noise, brings about many negative impacts [1-3]. One of the most important ways to reduce carbon dioxide emissions to get rid of oil dependency for human is the development of new energy vehicles and promoting low carbon transport [2]. Electric cars can be charged using a socket connected to the grid, large-scale grid connection for charging electric vehicles will pose huge impacts on the grid, especially if users charge electric vehicles in the same period, which will improve the maximum power load, aggravate the load peak and off-peak difference, increase the difficulty of controlling the power grid optimization, impact power quality, and reduce the life of distribution transformer. Electric vehicle charging load prediction can reduce the impact on power grid caused by the electric cars connected to the grid, and provide reference for optimizing operation and planning of power grid. Traditional power load methods of prediction, such as time series prediction method, wavelet analysis, chaos theory, and the neural network, are passive predictions that depend on

---

<sup>\*</sup> Fund project: the national natural science foundation of China(51277057); Research of load forecasting theory and method under smart grid environment.

<sup>\*\*</sup> Corresponding author.

a large number of historical data and related influence factors . However, the electric cars industry is an emerging technology in China, there is few actual historical data of electric cars on the road. The prediction for EVS charging load mostly depends on the drive behavior and the statistics analysis of traditional fuel vehicle.

Electric vehicle charging demand of statistical model is established in [4], according to the statistical data of fuel vehicles, combined with the influencing factors of electric vehicle charging load. But EVS are supposed to charge only at home, and will begin to recharge immediately as soon as the cars get home. In [5] the charging load prediction is according to the number of electric cars of entering and leaving in a certain area for a period of time. It also assumes that electric vehicles recharge at home, and the charging power is constant, all travels outside can be supported with a single battery which is full charged. Paper [6] analyzed the important factors that affect the electric car charging such as parking location, travel distance, driving time, parking time , based on the driving mode of automobile, then attained the load profile in different areas and in different parking locations. Due to the electric car is still in development stage, there are a few charging facilities in public place, most of the electric cars will be charged at home [8]. Paper [7] traced and investigated a large number of the electric car drivers to study the driving behavior of electric cars, analyzed the probable charging location and charging time. The EVs recharged not only once a day or only at home in their actual use. Drivers decide to charge or not largely depends that if the remaining state of charge (SOC) can meet the power demand for next trip or whether the parking duration in given locations can meet the need of charging Fees of charging and parking, Income, oil prices, and the convenience of the charging infrastructure also will affect the users' decision to charge. But these factors are negligible compared to SOC and parking duration. The charging facilities will be improved much and increase gradually, because of the country's heavily promoted on electric cars. Therefore, the possibility of charging in residential, workplace and in the large-scale parking lot will be increased much, electric vehicles batteries can be charged at the main parking place successfully in the future.

This paper analyze the travel characteristics of electric vehicles, and then use the fuzzy inference system to emulate the process of drives deciding to charge their cars, the charging probability is attained in the given location. Finally, the daily profile of charging load can be predicted according to the numbers of electric vehicles forecasted in Beijing.

## 2 Travel Characteristics of EVS

The charging load of EVS has a strong randomness in temporal and space, which is determined by the driver behavior of the users when they use their cars. Travel features contain the information such as where they park their cars, parking duration, when they arrived at the destination and the state of charge when arrived. Therefore, understanding the electric car travel characteristics in time and space will be of great significance to forecast charging load distribution of the electric vehicles. Travel characteristics of vehicle are analyzed according to the Beijing municipal transportation development report in 2005 and the third trip survey [12, 13].

## 2.1 The Purpose of Travel

People usually drive their cars for commuting or not. The activities of commuter include the travel to school and work places to and from, the other non-commuting activities include shopping, entertainment and so on. Different travel purpose has different parking behavior, parking time will be last more than 3 hours in most of the commuter parking and residential parking, while the parking for business or other pleasure purpose the parking duration will be mostly concentrated in the 30 minutes between 3 hours. The time of parking in residential area are more concentrated, which usually last 10 hours, compared with the commuter parking [9]. The average travel daily times is 3.16 [13]. So the main travel purpose can be assumed to be at home, workplaces and large business places. Users generally go to work from residential areas in the morning, then go shopping or dining after the work, finally arrive home. In this paper, only the car users who have the fixed work will be discussed. The unemployed or retired, who's parking place and time is not fixed, and only a minority, are not discussed in this paper.

## 2.2 Residential Parking Time and Duration Analysis

The users generally set out from home in the morning in weekdays, and finally return home after a day of activity, and then the cars would stay at home for a long time till the next trip. Electric vehicles will have a great possibility to be recharged in the evening .Fig.1 shows the begin time of residential parking [14]. Only one peak of parking time in the residential area can be seen from the Fig.1, namely the evening peak at 17:00 - 19:00, the number of parking cars accounts for 51.09%, the cars generally stay at home till the next morning, parking duration in residential area generally last more than 10 hours, so the residential is often considered the most likely location to be charged for EVS.

## 2.3 Parking Time and Duration Analysis of Work Places

The drivers usually go to work at 7: 00-8: 00, which is during the rush hour, and most of the cars arrive at the work places at 8: 00-9: 00, accounted for 26%. The afternoon rush hour is between 17:00 and 18:00, and reaches the peak at 17:30, accounted for 35%. The peak time of arriving home is between 18:00 and 19:00 [12]. Because the work places are relatively fixed, the cars usually can stay for a long time, generally close to the working time 8 hours. The EVS can recharge when arrived at work places until their next trip after work. Due to the long parking duration at work places, the possibility of charging for EVS will be large.

## 2.4 Parking Time and Duration Analysis of Commercial Area

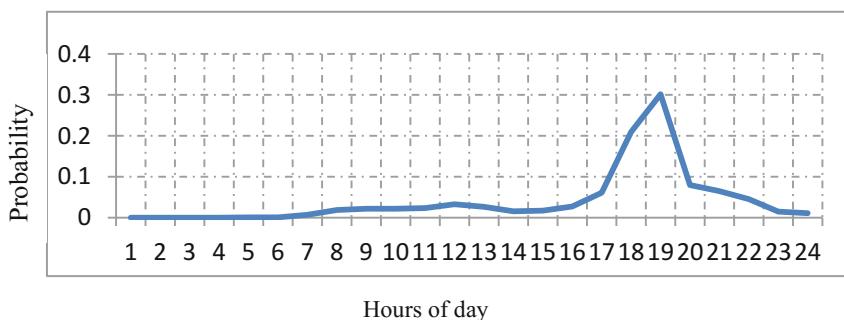
The main parking purpose in commercial area is commonly used to shopping, catering or entertainment. Parking time is short, most concentrated within 2 hours. The distribution of parking duration in commercial area is show in Fig.2. The average parking time is 79.8 minutes [10]. So if the charging infrastructure is available in shopping centers, the cars are most likely to be recharged in this place. Parking time in commercial places is usually at 12:00-13:00 and 17:30-19:00.

## 2.5 Travel Distance

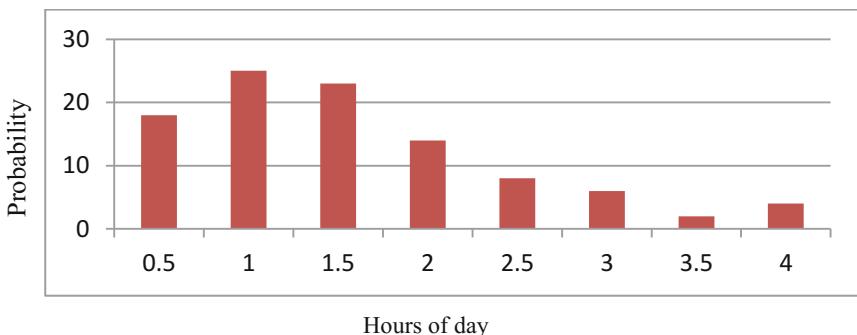
The power demand of electric cars determines the charging behavior, whether the electric cars are charged or not is related to the battery remaining power closely. The remaining state of charge in the electric vehicle battery is used to represent the charge demand in this paper. The charging time and the power energy drawn from the grid largely depend on the initial SOC before charging. SOC decreased linearly with the distance increasing. The average distance of private cars in Beijing is 19596 km, the average daily range of about 53.7 km, 14km per time [12]. BYD F3DM, for example, the power consumption per 100km is 16 kWh; the maximum range in electric mode is 60 km. The electric cars are likely to be charged after arriving at given destination. Travel mileage satisfy lognormal distribution according to the statistical analysis of the traditional car data, as shown in formula (1)

$$f_D(x) = \frac{1}{x\sigma_D\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu_D)^2}{2\sigma_D^2}\right] \quad (1)$$

Where  $x$  is the daily distance,  $\mu_D=3.2$ ,  $\sigma_D=0.88$ . As the cars traveling three times a day, the distance traveled to work can be assumed to be  $x / 3$ , the distance traveled to commercial area can be  $2x / 3$ .



**Fig. 1.** The time distribution of starting to park in residential area



**Fig. 2.** The parking duration (hours) in commercial area

### 3 The Fuzzy Inference System of the Process to Decide to Charge

The fuzzy inference system (FIS) is an advanced algorithm framework, and widely applied in automatic control, artificial intelligence, pattern recognition, and many other fields. In this paper, the fuzzy inference is applied to emulate the process of decision-making for a driver when deciding to charge the vehicle's battery. The input of fuzzy system is the distance before charging and the parking duration in a given location, then the inference rules are built according to users' charging habits, finally the charging probability as an output of fuzzy system can be attained.

The charging behavior of electric vehicles largely determines distribution of space and time of the charging load. Drivers decide to charge or not largely depends that if the remaining SOC can meet the power demand for next trip or whether the parking duration in given locations can meet the need of charging. If the SOC is too low, the next trip can't complete in electric mode for a plug-in hybrid. Drivers tend to charge the battery in the destination, if the parking duration is long, users also tend to charge the battery to full. Users are more willing to use power energy because of its saving and economic advantages and the price is low compared to oil price. Vague language will be used to describe the current SOC and the parking duration, just as charging a mobile phone, the actual accurate percentage of the remaining power will not be cared about. Users usually only estimate how much energy remained and how long the battery can be recharged fully. It is not easy to calculate SOC, but distance can be directly read from the odometer. It is easier to estimate the probably distance of next trip, and then decide whether to charge when knowing the maximum distance in electric mode for a plug-in hybrid. The language of fuzzy variables on distance traveled when arriving destination and the parking duration in the given place is used in this paper. Different electric cars have different battery capacity, so the maximum range in electric mode is different. Driving distance before charging is normalized based on its maximum all-electric range. The fuzzy set "low", "medium", "high" is used to cover the whole working area of the distance when arrived the destination. Fig.3 (a) shows membership functions created for the distance traveled before charging. Similarly, the length of time for parking also expressed by three linguistic terms (Short, Average, Long), as shown in Fig.3 (b). Parking events ranging from 30minutes to around 4 hours is labeled "M", less than 30minutes is labeled "S". Parking duration at home or work places is very long according to the above analysis. So the batteries of electric vehicles are likely to be recharged in these two regions. The reasoning rules of fuzzy inference rules are established according to the user's experience and the analysis of electric vehicle charging behavior, as shown in Table 1. Outputs produced by these rules acted on the two inputs of the fuzzy system, the membership functions created for charging probability is shown in Fig.3 (c). Then aggregated and defuzzified to gain the probability of charging in different location. Five linguistic terms is used to cover the whole area of the charging probability. The center-of-mass operation is used for defuzzification, more calculate details can be seen in [15-16].

## 4 Prediction and Analysis of Charging Load

### 4.1 Type of Battery

The type of battery determines the battery capacity, mileage, charging power characteristics. The rechargeable batteries commonly used include lead-acid battery, nickel cadmium battery, nickel metal hydride batteries and lithium ion batteries. Due to the extremely high performance advantages of lithium ion batteries, it will be the inevitable direction of power battery in the future. This article assumes that all electric vehicles use lithium ion batteries. Battery capacity designed can satisfy the demand of the day's mileage, but the volume is greater, the price will be higher accordingly. From the future perspective of the development, Cars can be likely to be recharged anywhere at any time .So while maintaining the same mileage, the battery capacity can be significantly reduced, thereby reducing the cost of production of electric vehicles, improving penetration of electric vehicles, especially for private cars travelling short. Lithium battery capacity in this article is assumed to be 16kWh.

### 4.2 Charging Mode

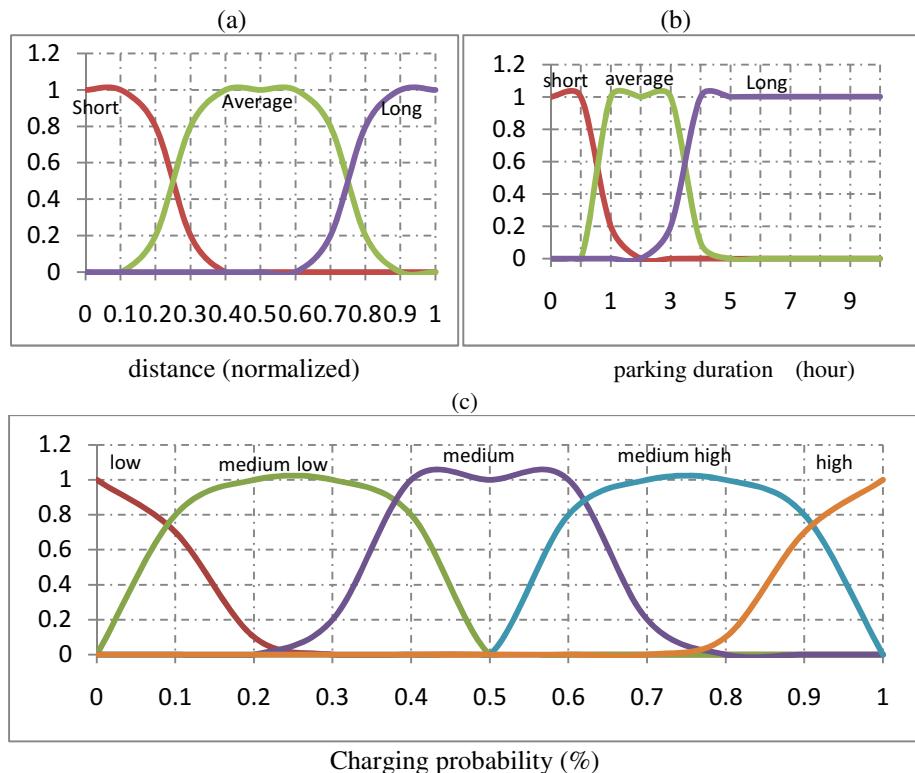
There are four ways of charging, according to China's automotive industry standards, as shown in Table 2.This article selects the charging way of 220 V (AC) / 16 A, namely the charging power is 3.52 kW. Charging mode determines the length of time for charging. The same stop time, but the charging mode is different, the intention of drivers' decision to charge will be different. Even the parking duration is short, if the charge mode is fast-charge, the possibility of charging for electric vehicles is still large, while the drivers would not recharge their cars if the charge mode is slow-charge with short parking duration. The membership function is based on the slow charge in this paper.

### 4.3 The Ownership of Electric Cars Prediction

The ownership of private cars in Beijing reached 4.075 million by 2012. The average annual growth rate of private vehicles is 14% several years ago, but it has slowed down significantly, because of the restriction measures on the cars in Beijing. The growth rate of private vehicles is only 3% in 2011.The policy for cars will not be changed much in Beijing for a long period of time. Therefore, the annual growth of private vehicles in Beijing is likely to remain 4%, so the total number of vehicles in 2015, 2020, 2030 can be predicted. The results shown as follows:

Year	2015	2020	2030
Number (ten thousand)	458.4	557.7	825.5

According to incomplete statistics of China Association of Automobile Manufacturers, the total new energy vehicles sold are 24000 in China from 2011 to the first half of 2013, while in Beijing are 3388, which are ranked the third in China, the number of private electric cars account for one percent of the total private vehicles in 2012. The number of electric vehicles in 2015, 2020, 2030 is assumed to account for 2%, 10%, and 30% of the total private cars in this paper.



**Fig. 3.** Membership functions. (a) normalized driving distance; (b) parking duration;(c) charging probability

**Table 1.** Rules of the fuzzy system

If distance is	And parking duration is	Then probability of charging
Low	Short	Low
Low	Average	Medium Low
Low	Long	Medium
Average	Short	Medium Low
Average	Average	Medium
Average	Long	Medium High
Long	Short	Medium Low
Long	Average	Medium High
Long	Long	High

**Table 2.** Standard of charging ways

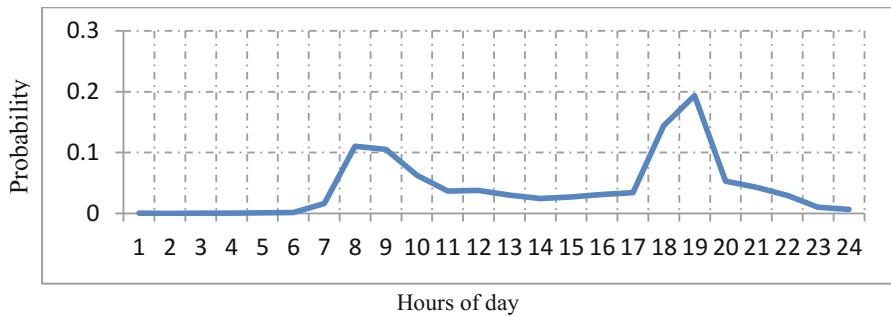
CHARGING MODE	RATE VOLTAGE	RATE CURRENT
1	220V(AC)	16A
2	220V(AC)	32A
3	380V(AC)	32A
4	400V/750V(DC)	125A
	400V/750V(DC)	250A

#### 4.4 Prediction Steps

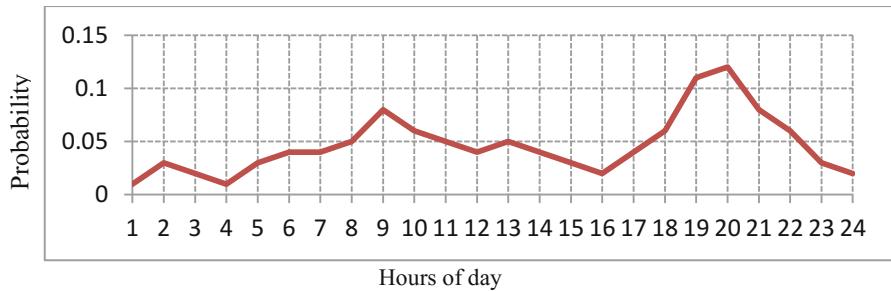
One day can be divided into three periods, namely the corresponding three charging period in the corresponding locations. When the forecast time is between 17:00-18:30, the electric cars are mostly concentrated in business areas to charge the battery using fast charging mode at this moment. Similarly, the time in residential area and working places can be assumed according to the time when the cars begin to be parked.

The probability of parking events in a day can be seen in Fig.4, according to 146950 parking records in Beijing [14]. The total number of cars parked in a fixed time can be attained according to the percentage in Fig.6 .The parking places can be known according to the forecasted time and the distance traveled before charging can be produced randomly by the daily probability density distribution. The length of time for parking and the distance traveled when arrived at the destination is used as the inputs of fuzzy system, and then the reasoning rules are activated. Finally, the hourly charging probability profile will be gained by defuzzification, as shown in Fig.5. Assuming that all vehicles will be recharged until the battery is full or till to the end of the time of parking. If the charging of electric vehicles will continue to the next moment, the charging probability at the next moment is 1. As shown in Fig.6, the parking events probability is 10.54% at 9 a.m., the number of electric vehicles is 11600 in 2015 according to the prediction above. Then, there will be  $11600 * 10.54\% = 1122$  vehicles parking at this moment. The charging location can be supposed to be working places. The charging probability is 0.08 at 9 a.m. as shown in Fig.5. The charging power is  $220*16=3.52$ kW, so the total charging load is  $3.52 * 1122 * 0.08 = 344.1$  kW. Because not all the electric cars will travel, assuming that 80% of the electric cars will travel and will be recharged in the parking locations. A random number r is produced, when  $r <= 0.8$ , the electric cars can be assumed to be recharged.

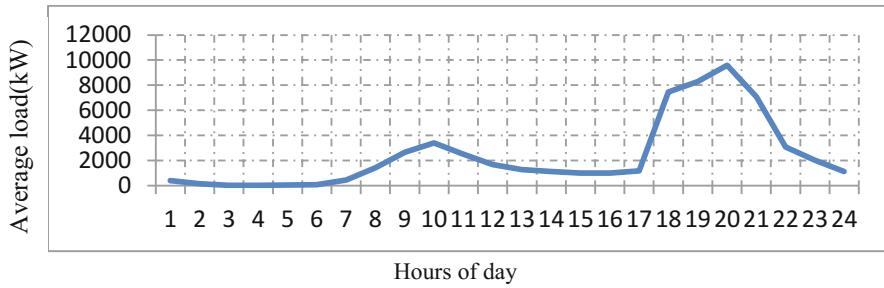
Assuming that the charging efficiency is 90%, the lithium battery charging power is 3.52 kW, and 12.26 kW. The time of day can be divided into 24 points. The expect of total charging power of all the electric vehicles parked in the specific time can be gained by Monte Carlo simulation, then the daily profile of charging load is attained.



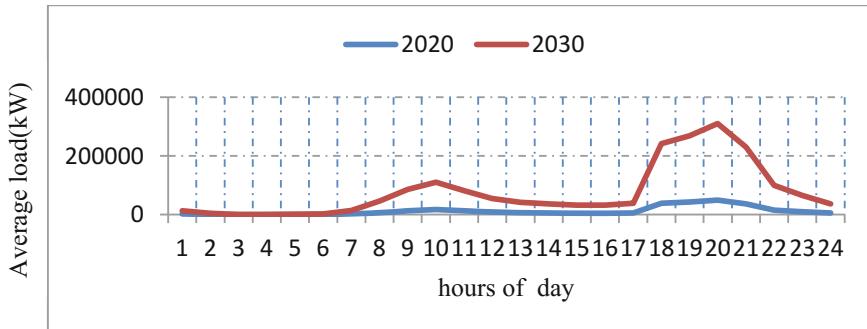
**Fig. 4.** The probability of parking in weekdays



**Fig. 5.** The average probability of charging in weekdays



**Fig. 6.** Average load demand in weekdays in 2015



**Fig. 7.** Average load demand in weekdays in 2020 and 2030

## 5 Conclusion

The predicted profile as shown in Fig.6 and Fig.7 showed two peaks obviously, maximum load appears at night, because most of the private cars tend to charge their cars at night. The maximum load is respectively 9576.0kW, 49799kW, and 310745.9kW in 2015, 2020, 2030. The charging load is increasing dramatically due to the development of EVS. The results show that the time of peak is almost the same as the regular load peak, which will pose huge impacts on power grid. The profile of charging load is significant to study the power quality and load voltage loss. It also has reference meaning for studying the optimization of charging and vehicle to grid.

## References

1. Zhang, W., Wu, B., Li, W., Lai, X.: Discussion on Development Trend of Battery Electric Vehicles in China and Its Energy Supply Mode. *Power System Technology* 33(4), 1–5 (2009)
2. Xin, H.: The development of low carbon economy and electric vehicles: trends and counter measures. *China Opening Herald* (5), 31–35 (2009)
3. Hu, Z., Song, Y., Xu, Z.: Impacts and Utilization of Electric Vehicles Integration Into Power Systems. *Proceedings of the CSEE* 32(4), 1–10 (2012)
4. Tian, L., Shi, S., Jia, Z.: A Statistical Model for Charging Power Demand of Electric Vehicles. *Power System Technology* 11, 126–130 (2010)
5. Tikka, V., Lassila, J., Haakana, J., Partanen, J. (eds.): Case study of the effects of electric vehicle charging on grid loads in an urban area. In: 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe). IEEE (2011)
6. Tang, D., Wang, P. (eds.): Dynamic electric vehicle charging load modeling: From perspective of transportation. In: 4th IEEE/PES Innovative Smart Grid Technologies Europe (ISGT EUROPE). IEEE (2013)
7. Adornato, B., Patil, R., Filipi, Z., Bareket, Z., Gordon, T.: Characterizing naturalistic driving patterns for Plug-in Hybrid Electric Vehicle analysis. In: IEEE VPPC 2009, Ann Arbor, MI, pp. 655–660 (2009)
8. ECO tality, “The EV project,” Q1-2012 Report
9. Pan, C., Zhao, S.: The analysis of different parking behavior with different trip purpose. *Construction Economy* 1 (2011)
10. Luo, Z., Hu, Z., Song, Y.: Calculation method of EV charging load. *Automation of Electric Power System* 14, 36–42 (2011)
11. Lee, T.K., Adornato, B., Filipi, Z.S.: Synthesis of real-world driving cycles and their use for estimating PHEV energy consumption and charging opportunities: Case study for Midwest/US. *IEEE Transactions on Vehicular Technology* 60(9), 4153–4163 (2011)
12. Beijing Traffic Development Report in 2005
13. The third travel survey of Beijing residents
14. Li, Z., Zong, F., Zhang, B.: Study on Parking Information Behavior in Large and Medium-sized Cities. *Information Science* 29(12), 1896–1901 (2012)
15. Wang, Q.: The Research on Fuzzy Inference of Online Consumers’ Buying Intention. Dalian University of Technology (2013)
16. Liu, H.: Mamdani fuzzy reasoning algorithm and its Matlab realization. *The Science Education Article Collects* 10, 269–271 (2008)

# Security Event Classification Method for Fiber-optic Perimeter Security System Based on Optimized Incremental Support Vector Machine

Lu Liu<sup>1</sup>, Wei Sun<sup>1</sup>, Yan Zhou<sup>1</sup>, Yuan Li<sup>2</sup>, Jun Zheng<sup>2</sup>, and Botao Ren<sup>2</sup>

<sup>1</sup> PetroChina Pipeline R&D Center, Langfang, Hebei, China

<sup>2</sup> PetroChina Pipeline Company, Langfang, Hebei, China

lordman1982@163.com

**Abstract.** The way of efficiently classifying the fence climbing, fabric cutting, wall breaking and other environment factors, is an imperative problem for fiber-optic perimeter security system. To solve this problem, a security threats classification method based on optimized incremental support vector machine is proposed. In this method the artificial bee colony algorithm is introduced to optimize the penalty factor and kernel parameter of incremental support vector machine under specified fitness function, and the optimized incremental support vector machine is used to classify the perimeter security threats. To testify the performance of the proposed method, the experiment based on UCI datasets and actual vibration signal are made. Comparing with the support vector machine optimized by other algorithms, higher classification accuracy and less time consumption is achieved by the proposed method. Therefore, the effectiveness and the engineering application value of this proposed method is testified.

**Keywords:** Perimeter Security, Incremental Support Vector Machine, Artificial Bee Colony Algorithm, Parameter Optimization, Wavelet Transform.

## 1 Introduction

There is no doubt that the station of oil pipeline plays an important role in oil transportation. Because of the increasingly grim situation of terrorism in China, new perimeter security systems based on pattern recognition are implemented for the security of oil station. In perimeter security system, a key problem is how to efficiently classify the true alarms caused by fence climbing, fabric cutting, wall breaking, and nuisance alarms cause by environment factors such as small animal and wind.

Support vector machine(SVM) is a popular pattern recognition method based on small sample learning. Because of its many advantages, SVM has been widely applied in the research fields of security event classification [1,2,3]. Incremental support vector machine(ISVM) is an improvement of basic SVM. New support vectors can be extracted from new introduced training samples, and the classification model can be adjusted continuously. In the classification process with ISVM, parameter selecting has a critical influence on the final classification accuracy. The common parameter optimization methods include: gradient descent algorithm(GD)[4], simu-

lated annealing algorithm(SA)[5,6], ant colony optimization algorithm(ACO)[7,8], genetic algorithm(GA)[9,10] and particle swarm optimization algorithm(PSO)[11,12]. It has been proved that the function relationship between the parameters and the classification accuracy of ISVM contains many local peaks, and it is dissatisfying that the optimization methods above would lead into the local optimal solution in the search process[13,14]. In the year of 2005, a new optimization algorithm-artificial bee colony(ABC) algorithm, which is inspired by the foraging behavior of bee colony, is proposed by Karaboga[15]. In the ABC algorithm, exploration in new search domain is carried out in parallel with exploitation in the known domain, so the problem of leading into local optimal solution can be avoided, and the performance of this algorithm is better than other optimization algorithms[14,15,16].

In this paper, a novel security event classification method based on optimized incremental support vector machine is proposed. Fence climbing, fabric cutting, wall breaking and wind are treated as four kinds of typical alarm-triggering events. The ISVM with RBF kernel function is used as the classifier, and ABC algorithm is utilized to optimize the parameters of ISVM. Then the optimized ISVM is used to classify security events. The comparison experiments among the proposed method and the ISVMs optimized by different optimization algorithms, which is through UCI feature dataset and actual signal, are made to testify its performance of classification accuracy and time consumption. The detail of this proposed method is described in the following sections.

## 2 Parameter Optimization of Support Vector Machine

### 2.1 Artificial Bee Colony Algorithm Principle

In ABC algorithm, three groups of bees are contained in the colony of artificial bees: employed bees, onlookers and scouts. Each search cycle consists of three steps: moving the employed and onlooker bees onto the food sources and calculating their nectar amounts and determining the scout bees and then moving them randomly onto the possible food sources. A food source, whose amount is SN, represents a possible solution to the problem to be optimized. The nectar amount of a food source corresponds to the quality of the solution. Onlookers are placed on the foods by using 'roulette wheel selection' method. Every bee colony has scouts that are the colony's explorers. In ABC algorithm, one of the employed bees is selected and classified as the scout bee. The classification is controlled by a control parameter called 'limit'. If a solution representing a food source is not improved by a predetermined number of trials, then that food source is abandoned by its employed bee and the employed bee associated with that food source becomes a scout. The number of trials for releasing a food source is equal to the value of 'limit', which is an important control parameter of ABC algorithm.

The main steps of the ABC algorithm program are given follows:

```
Initialize
REPEAT
```

- Move the employed bees onto their food sources and determine their nectar amounts.

- Move the onlookers onto the food sources and determine their nectar amounts.
  - Move the scouts for searching new food sources.
  - Memorize the best food source found so far.
- UNTIL (requirements are met)

## 2.2 The Optimization Process of Support Vector Machine

In the proposed method, the ISVM parameters need to be optimized include: the penalty factor  $C$  and kernel parameter  $\sigma$ . The initial setting of the proposed method is as follows:

- Initializing the parameters of ABC algorithm, which include food source  $S_N$ , the max iteration time of single food source  $limit$ , and termination iteration time of whole search cycle  $N_{mc}$ .
- Specifying the fitness function of ABC algorithm. Considering the main purpose of the optimization is achieving high classification accuracy, the formula(1) is specified as the fitness function:

$$V_{obj} = 1 - V_{acc} \quad (1)$$

$V_{acc}$  is the classification accuracy.

- Determining the search range of the target parameters.

In each search cycle, the ABC algorithm is used to obtain a solution of the ISVM parameters, and the feature vectors of the training samples are put into the ISVM for training. After then, the trained ISVM is used to classify the test samples, and classification accuracy is obtained. This accuracy is put into the fitness function to estimate the solution's quality. After judgement the solution is accepted or abandoned according to ABC algorithm's principle, then the next iteration of search cycle is started. The search cycle doesn't stop until the termination iteration time of whole search cycle is met.

The experiments in the references[14,15,16] have proved that ABC algorithm possesses better optimization ability than other optimization algorithms. To testify the performance of the proposed method, the experiment based on the UCI standard dataset[17] is made in this paper. The datasets are split into training samples and test samples randomly. In this experiment, the proposed method is compared with the ISVMs optimized by ant colony optimization algorithm(ACO), genetic algorithm(GE) and particle swarm optimization algorithm(PSO) through the representative dataset of Heart, Iris, Wine and Glass from UCI dataset, which are shown in table 1. After repeated tests, the food source number  $S_N$  is set to 20, the max iteration time of single food source  $limit$  is 50, and termination iteration time of whole search cycle  $N_{mc}$  is 1000. The parameters' search range is [0.1, 1000]. The classification accuracy of every method is given in table 2. It is concluded from the accuracy result that, because of the parallel exploration and exploitation mode, the proposed method can achieve higher classification accuracy. The time cost of every method is given in table 3. They are the average values of 3 times repeated experiments. From table 3, it is revealed that all the four methods' time cost increase as the dimension and class number of the dataset

increase. Under the condition of big class number, the proposed method's time cost is more than other methods. However, under the condition of small class number, the proposed method shows better convergence, and cost less search time than other methods.

**Table 1.** Instruction of UCI dataset

Dataset (class number)	Dimension	Training Sample	Test Sample
Heart(2)	13	190	86
Iris(3)	4	90	60
Wine(3)	13	105	65
Glass(6)	9	99	95

**Table 2.** Accuracy result of experiment

Dataset (class number)	GA-ISVM	ACO-ISVM	PSO-ISVM	ABC-ISVM	C	$\sigma$
Heart(2)	78.5	80.1	79.3	84.4	81.2	2.7
Iris(3)	96.7	97.3	93.3	99.0	200.9	2.1
Wine(3)	93.6	91.9	89.9	95.8	205.1	2.2
Glass(6)	64.4	61.7	65.0	64.0	38.2	0.3

**Table 3.** Time cost result of experiment

Dataset (class number)	GA-ISVM	ACO-ISVM	PSO-ISVM	ABC-ISVM
Heart(2)	172.1	189.2	177.7	160.5
Iris(3)	96.4	92.9	88.5	86.1
Wine(3)	115.2	110.1	129.4	109.3
Glass(6)	141.5	139.7	121.7	145.3

### 3 Experiment Study

#### 3.1 Data Acquisition and Noise Reduction

Actual signal experiment is made to testify the performance of the proposed method. The actual vibration signal for the experiment is acquired by fiber-optic perimeter security system on Zhengzhou and Daqing oil pipeline station. Fence climbing, fabric cutting, wall breaking(using electric drill) and wind are treated as four kinds of typical alarm-triggering events, and all of them are carried out in 5 different positions of each security zone(20 zones total). For each kind of these alarm-triggering events, 100 group signal samples are acquired separately. Half of them are treated as training samples, and the other half are treated as test samples.

Before the alarm-triggering events classification by the proposed method, noise reduction and feature extraction must be implemented. After observation, it is found that the original signal is composed of narrowband useful signal and broadband noise. Therefore, a noise reduction method based on one order Wiener filtering, which is usually used in speech signal processing, is utilized in this method.

### 3.2 Feature Extraction

In the step of feature extraction, the denoised signal is analyzed through time-frequency analysis method, and the features which can reflect the characteristic of signals are extracted and combined into feature vectors. It is found that the signal energy distributions in the frequency domain of these four kinds of threat events are different from each other, therefore the energy of frequency bands is chosen as the main features of the signal. Firstly, wavelet transform is utilized to decompose the denoised signal into several frequency bands, the wavelet basis is db3, and the layer number of wavelet is 8. After transformation, 1 group of approximate wavelet coefficients and 8 groups of detail wavelet coefficients are acquired. The energy of every frequency band is calculated as follows:

$$E(i) = \frac{1}{N-1} \sum_{t=1}^N (f_i(t))^2 \quad (2)$$

$f_i(t)$  is the  $t$ th wavelet coefficient of the  $i$ th layer, and  $N$  is the amount of the coefficients in every layer.

Then the wavelet energy is normalized by this formula:

$$E(k) = \frac{E(k)}{\sum_{i=1}^9 E(i)} \times 100 \quad (3)$$

The main feature vector is combined by the normalized wavelet energy of every frequency band.

Considering the similarity of energy distribution between the fence climbing signal and wind signal, another feature is needed to distinguish the two kinds of signal. After observing the time-domain signal, it is found that the time length beyond the threshold could be considered as the assistant feature. At last, the frequency-domain features and the time-domain features are combined into the final feature vectors, as shown in table 4.

### 3.3 Optimization and Classification

After noise reduction and feature extraction, the proposed method is utilized to achieve global optimal solution of ISVM parameters and high classification accuracy of security events. After repeated tests, the food source number  $S_N$  is set to 35, the max iteration time of single food source  $limit$  is 100, and termination iteration time of whole search cycle  $N_{mc}$  is 400. The parameters' search range is [0.1, 1000].

**Table 4.** Feature vectors of the security events

Security event	Feature vector									
Fence climbing	20.63	16.42	15.55	12.21	8.99	8.21	7.83	6.81	3.35	15
Fence climbing	21.69	17.73	15.11	10.24	8.98	8.09	8.03	5.66	4.47	17
Fence climbing	20.01	17.42	16.20	10.73	9.33	7.65	8.29	5.01	5.36	18
Fabric cutting	26.54	22.16	8.88	10.01	7.64	7.04	8.14	5.21	4.38	3
Fabric cutting	24.98	23.26	9.15	10.18	8.06	6.89	7.22	5.17	5.09	7
Fabric cutting	24.70	24.15	9.77	12.21	7.94	7.26	5.81	4.01	4.15	5
Wall breaking	7.08	8.09	8.92	44.21	6.14	7.26	8.15	6.36	3.79	23
Wall breaking	6.21	9.49	8.03	42.54	7.36	8.99	8.35	5.36	3.67	20
Wall breaking	6.99	10.04	6.91	42.68	6.26	8.42	7.29	6.33	5.08	21
Wind	17.03	21.98	16.38	10.11	9.22	9.47	6.99	4.35	4.47	20
Wind	17.84	20.02	15.81	10.97	9.47	9.49	7.91	3.87	4.62	22
Wind	18.01	21.38	16.55	9.41	8.56	10.33	7.47	4.09	4.20	22

To testify the proposed method's performance, the proposed algorithm is compared with the ISVMs optimized by ant colony optimization algorithm, genetic algorithm and particle swarm optimization algorithm. The experiment's accuracy results are shown in table 5-8. As shown in the tables, the classification accuracy of ACO-ISVM, GA-ISVM and PSO-ISVM are 88.5%, 81%, 85.5%. The total accuracy of the proposed method is 92%, and particularly the accuracy towards the wall breaking is 100%. The final optimal solution of penalty factor  $C$  and kernel parameter  $\sigma$  is 389.2 and 47.4. It is testified that, as a result of the advantage of ABC algorithm, the local and global optimal solutions are both given consideration to, and high classification accuracy can be achieved through the parallel exploration and exploitation for the parameters in a finite domain. The time consumption is also tested, and the results are shown in table 9. Every method is tested 3 times and the average time consumption is calculated. It is shown that the proposed method possesses less time consumption. Therefore, it can be concluded from the experiment results that the proposed method is very effective in pipeline station security events recognition.

**Table 5.** Test result of ACO-ISVM

Fact\Experiment result	Fence climbing	Fabric cutting	Wall breaking	Wind
Fence climbing	45	1	0	4
Fabric cutting	4	40	1	5
Wall breaking	0	2	46	2
Wind	4	0	0	46

**Table 6.** Test result of GA-ISVM

Fact\Experiment result	Fence climbing	Fabric cutting	Wall breaking	Wind
Fence climbing	39	4	0	7
Fabric cutting	8	34	0	8
Wall breaking	0	3	45	2
Wind	6	0	0	44

**Table 7.** Test result of PSO-ISVM

Fact\Experiment result	Fence climbing	Fabric cutting	Wall breaking	Wind
Fence climbing	42	2	0	6
Fabric cutting	6	35	0	9
Wall breaking	0	0	47	3
Wind	3	0	0	47

**Table 8.** Test result of ABC-ISVM

Fact\Experiment result	Fence climbing	Fabric cutting	Wall breaking	Wind
Fence climbing	44	2	0	4
Fabric cutting	8	42	0	0
Wall breaking	0	0	50	0
Wind	2	0	0	48

**Table 9.** Time consumption of methods

ACO-ISVM	GA-ISVM	PSO-ISVM	ABC-ISVM
147.1	110.3	104.3	99.8

## 4 Conclusion

A novel security event recognition method based on optimized ISVM has been proposed in this paper. To solve the problem of ISVM's parameter selection, ABC algorithm is introduced as the optimizer. In the proposed method, the ABC algorithm is utilized to optimize the penalty factor  $C$  and RBF kernel parameter  $\sigma$  under the specified fitness function in a finite domain, and the optimized ISVM is used to classify the feature vectors. Through the experiments under the UCI dataset, because the exploration and exploitation are both given consideration to in the parameter search process, it is proved that the proposed method can achieve higher classification accuracy in all

cases and cost less time under a small number of class. To prove its engineering value, the proposed method is also experimented in the perimeter security system used in the oil pipeline stations in China. Through the contrast experiments with ISVMs optimized by other algorithms, it is proved that the proposed method can achieve higher classification accuracy of the threat events with less time consumption. Therefore, the effectiveness and engineering application value of the proposed method is testified.

## Reference

1. Mehmood, A., Patel, V.M., Damarla, T.: Discrimination of Bipeds from Quadrupeds Using Seismic Footstep Signatures. In: International Conference on Geoscience and Remote Sensing Symposium, IGARSS2010, pp. 6920–6923. IEEE Press, Munich (2012)
2. Hu, Y., Lixin, L., Fangchun, D., Jin, H.: ANN-based Multi Classifier for Identification of Perimeter Events. In: 4th International Symposium on Computational Intelligence and Design, ISCID2011, pp. 158–161. IEEE Press, Hangzhou (2011)
3. Hu, Y., Guangshun, S., Qinren, W., Shangqin, H.: Identification of Damaging Activities for Perimeter Security. In: 1st International Conference on Signal Processing Systems, ICSS2009, pp. 162–166. IEEE Press, Singapore (2009)
4. Tseng, P., Yun, S.: A Coordinate Gradient Descent Method for Linearly Constrained Smooth Optimization and Support Vector Machines Training. Computational Optimization and Applications 47(2), 179–206 (2010)
5. Lin, S., Lee, Z., Chen, S., Tseng, T.: Parameter Determination of Support Vector Machine and Feature Selection Using Simulated Annealing Approach. Applied Soft Computing 8(4), 1505–1512 (2008)
6. Xu, Z., Zhao, Y., Wen, X.: State Prediction of Slagging on Coal-fired Boilers based on Simulated Annealing Algorithms and Support Vector Machine. East China Electric Power 39(3), 463–467 (2011) (in Chinese)
7. Alwan, H.B., Kumahamud, K.R.: Optimizing Support Vector Machine Parameters Using Continuous Ant Colony Optimization. In: 7th International Conference on Computing and Convergence Technology, ICCCT 2012, pp. 164–169. IEEE Press, New Jersey (2012)
8. Gao, F., Pu, H., Zhai, Y., Chen, L.: Application of Support Vector Machine and Ant Colony Algorithm in Optimization of Coal Ash Fusion Temperature. In: 2011 International Conference on Machine Learning and Cybernetics, ICMLC 2011, pp. 666–672. IEEE Press, New Jersey (2011)
9. Batsaikhan, O., Ho, C.K., Singh, Y.P.: A Genetic Algorithm-based Multi-class Support Vector Machine for Mongolian Character Recognition. Journal of Computer Science 8(1), 84–95 (2008)
10. Long, G.: GDP Prediction by Support Vector Machine Trained with Genetic Algorithm. In: 2nd International Conference on Signal Processing Systems, ICSPS2010, pp. V3-1-V3-3. IEEE Press, New Jersey (2010)
11. Wang, J., Zhang, Z., Zhang, W.: Support Vector Machine based on Double-population Particle Swarm Optimization. Journal of Convergence Information Technology 8(9), 898–905 (2013)
12. Huang, Q.: Fuzzy Support Vector Machine Using Particle Swarm Optimization for High-tech Enterprises Financing Risk Assessment. In: 2013 International Conference on Computational and Information Sciences, ICCIS2013, pp. 670–673. IEEE Press, New Jersey (2013)

13. Liu, C., Wang, X., Pan, F.: Parameters Selection and Stimulation of Support Vector Machines based on Ant Colony Optimization Algorithm. *Journal of Central South University: Science and Technology* 39(6), 1309–1313 (2008) (in Chinese)
14. Karaboga, D., Basturk, B.: A Powerful and Efficient Algorithm for Numerical Function Optimization: Artificial Bee Colony(ABC) Algorithm. *Journal of Global Optimization* 39(3), 459–471 (2007)
15. Karaboga, D.: An Idea based on Honey Bee Swarm for Numerical Optimization. Technical Report, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
16. Karaboga, D., Basturk, B.: A Comparative Study of Artificial Bee Colony Algorithm. *Applied Mathematics and Computation* 214(1), 108–132 (2009)
17. UCI Machine Learning Repository, <http://archive.ics.uci.edu/ml> (accessed July 21, 2009)

# Author Index

- Abdelkrim, Akbi II-117  
Abulikemu, Aireti I-411  
An, Ni II-138  
Aysa, Alimjan II-491
- Bai, Lu I-168  
Bai, Xiang I-391  
Bao, Feilong II-464
- Cai, Xiaoxu I-401  
Cao, Yijia I-44  
Chen, Bo I-229, II-351  
Chen, Fei II-97  
Chen, Feifei I-305  
Chen, Feng I-335  
Chen, Jinhui I-140  
Chen, Li II-12, II-72, II-455  
Chen, MoHan II-499  
Chen, Shengyong II-128  
Chen, Shenyong I-345  
Chen, Shuixian I-325  
Chen, Su-Shing I-229  
Chen, Wen-Sheng II-351
- Chen, Xi I-273  
Chen, Yu II-341  
Cheng, Fang II-107  
Cheng, Hong II-331  
Cheng, Long I-21  
Cheng, Xingting II-22  
Cong, Yang I-355  
Cui, Zhi I-253  
Cui, Zhichao I-293
- Ding, Jianwei II-63, II-218  
Dong, Junyu I-401  
Dongfeng, Cai I-101  
Dou, Yong I-81, II-228  
Du, Haishun I-196  
Duan, Wuhui I-159
- Fan, Wei II-455  
Fang, Bin I-436  
Fang, Leyuan I-151, I-159  
Feifei, Sun I-237  
Feng, Jun I-229
- Fu, Hao I-371  
Fu, Keren I-283  
Fu, Peng II-189  
Fu, Zhongliang I-243
- Gan, Haitao I-273  
Gang, ZhenXiao I-418  
Gao, Changxin I-305, II-444  
Gao, Guanglai II-464  
Gao, Hao I-263  
Gao, Hongxia I-363  
Gao, Yicheng I-140  
Ge, Shiming I-325  
Gong, Qianhui I-63  
Gong, Wenlong I-63  
Gu, Yanchun I-31  
Guan, Qiu II-128  
Guo, Fan II-169  
Guo, Hao I-426  
Guo, Lei I-81  
Guo, Longyuan II-255  
Guo, Wei II-63
- Hamdulla, Askar II-474  
He, Hongying II-575  
He, Xiaobo II-159  
He, Xiaoyan II-149, II-179  
He, Yonghao II-1  
Hou, Yandong I-196  
Hu, Qingpu I-196  
Hu, Shiyu II-585  
Hu, Wenzheng I-11  
Hu, Ying II-235  
Hu, Yueming I-363  
Hu, Zhen I-11  
Hua, Yajing II-149  
Huang, Chun II-508, II-555, II-565  
Huang, Hong I-210  
Huang, Linlin II-518  
Huang, Qingxiu II-565  
Huang, Xiaotong II-72  
Huang, Yan I-283  
Huang, Yawei II-22  
Huang, Yongming II-436

- Huang, Yongzhen II-81, II-218  
 Huang, Zhiwei II-107  
 Huaxiang, Zhang I-237  
 Ibrayim, Mayire II-474  
 Jia, Guimin II-303  
 Jiang, Cheng II-159  
 Jiang, Jiang I-273  
 Jiang, Wei II-209  
 Jin, Haiqiang I-445  
 Jin, Yiting II-199  
 Jing, Pengjie II-528  
 Jing, Wen II-499  
 Kang, Xudong II-89  
 Kuang, Xiaoqin I-305, II-444  
 Leng, Hua I-91  
 Li, Canbing I-44  
 Li, Chunlong II-53  
 Li, Dedi II-536  
 Li, Denggang II-44  
 Li, Huali II-44  
 Li, Jiachang I-183  
 Li, Jiancheng I-168  
 Li, Kaihan I-21  
 Li, Lijuan I-44  
 Li, Liu I-237  
 Li, Ruimei II-303  
 Li, Shijie I-81  
 Li, Shuangshuang II-1  
 Li, Shutao I-151, I-159, II-44, II-89, II-97, II-276  
 Li, Song I-426  
 Li, Tangbing II-189  
 Li, Weibo I-455  
 Li, Weisheng II-296  
 Li, Xiaolong II-392  
 Li, Xiaotang II-402  
 Li, Xinran II-22  
 Li, Xinzhaoy I-293  
 Li, Yanan II-303  
 Li, Yuan II-595  
 Li, Yuchong I-381  
 Li, Yue II-436  
 Li, Yulian I-355  
 Liang, Shan II-209  
 Liao, Qingmin II-536  
 Liao, Wei II-189  
 Lin, Hui II-89  
 Ling, Wang I-130  
 Ling, Zhigang I-168  
 Liu, Chenglin I-1  
 Liu, Dan I-151  
 Liu, Dijun II-32  
 Liu, Gan II-128  
 Liu, Guohai II-286  
 Liu, Huaping II-245  
 Liu, Jiamin I-210  
 Liu, Jiancong I-183  
 Liu, Jun I-401  
 Liu, Lu II-595  
 Liu, Min II-382  
 Liu, Qingjie I-314, II-117  
 Liu, Sheng I-445, II-199  
 Liu, Shuping II-321, II-372  
 Liu, Wenju II-209, II-428  
 Liu, Xiaolong I-1  
 Liu, Xiaoyan II-402  
 Liu, Yu I-335, II-321, II-372  
 Liu, Yuehu I-293  
 Liu, Yunhui II-245  
 Liu, Yunlong I-111  
 Liu, Zewen II-149  
 Liu, Zhe II-360  
 Lu, Wenjun I-63  
 Luan, Xiao II-296  
 Luo, Diansheng II-575, II-585  
 Luo, Fulin I-210  
 Luo, Lei I-140  
 Luo, Min II-508  
 Ma, Jie I-21  
 Ma, Yi II-138  
 Ma, Zezhong I-210  
 Ma, Zhengming I-31  
 Mao, Jianxu II-149, II-179  
 Miao, Zhenjiang I-426  
 Niu, Guo I-31  
 Niu, Xin I-81, II-228  
 Osman, Abdurusul I-411  
 Pan, Binbin II-351  
 Pan, Chunhong II-1  
 Pan, Haixia II-53  
 Peng, Hui II-169  
 Peng, Jinjin II-303

- Peng, Minfang I-91  
 Peng, Yishu I-219
- Qi, Mingjun I-91  
 Qian, Jianjun II-341  
 Qian, Qifeng I-263
- Rauf, Muhammad II-81  
 Ren, Botao II-595  
 Ren, Dongwei II-12  
 Ren, JiaXin II-499  
 Ruan, Yu II-32
- Sang, Nong I-273, I-305, II-444  
 Satoshi, Naoi II-455  
 Shang, Shuangyin I-371  
 Shang, Zhaowei I-436  
 Shen, Hongbin II-409, II-419, II-528  
 Shen, Jifeng II-286  
 Shen, Linlin II-311  
 Shen, Wei I-391  
 Sheng, Biyun II-266  
 Shi, Zhe I-464  
 Simayi, Wujiahemaiti II-474  
 Song, Li II-481  
 Song, Yuqing II-360  
 Su, Xiangdong II-464  
 Su, Yi I-91  
 Suen, C.Y. I-436  
 Sun, Bin II-97, II-276  
 Sun, Changyin II-266  
 Sun, Daren II-536  
 Sun, Fuchun II-245  
 Sun, Haiping II-409  
 Sun, Jun II-97, II-455  
 Sun, Shiliang I-54, I-120  
 Sun, Wei I-46, II-595  
 Sun, Xia I-229
- Tai, Ying II-341  
 Tan, Hu I-91  
 Tan, Jianhao II-392  
 Tan, Mengxia II-255  
 Tan, Shaojie II-22  
 Tan, Yi I-44  
 Tan, Yingwei II-209, II-428  
 Tang, Jin II-169  
 Tang, Yandong I-355  
 Tang, Yu I-168  
 Tang, Yuanyan I-436
- Tang, Yunqi II-63, II-218  
 Tang, Zheng II-360  
 Tian, Huawei II-218  
 Tian, Jing II-72  
 Turki, Turghunjan Abdukirim I-411  
 Tursun, Dilmurat II-474
- Ubul, Kurban II-491
- Wang, Baoyun I-263, I-464  
 Wang, Changping I-130  
 Wang, Fan I-455  
 Wang, Huafeng II-53  
 Wang, Jiarong I-229  
 Wang, Liang II-81  
 Wang, Lili I-243  
 Wang, Lingfeng II-1  
 Wang, Ming II-22  
 Wang, Peiyan I-101  
 Wang, Runmin I-305, II-444  
 Wang, Weihua II-464  
 Wang, Weining I-183  
 Wang, Xiangyu I-335  
 Wang, Xiaoyan II-128  
 Wang, Xinggang I-391  
 Wang, Xinkun II-575  
 Wang, Xueting I-63  
 Wang, Yunhong I-314  
 Wang, Yuzhuo II-331  
 Wang, Zengfu II-321, II-372  
 Wei, Chongyang I-371  
 Wei, Hongxi II-464  
 Wu, Ao II-436  
 Wu, Chunpeng II-455  
 Wu, Fangrong I-91  
 Wu, Fuchao II-545  
 Wu, Jianhui II-255  
 Wu, Lin I-314  
 Wu, Qiang I-283  
 Wu, Tao I-371  
 Wu, Yefan II-189  
 Wu, Yingjing II-235
- Xiang, Jun II-444  
 Xiang, Peng II-382  
 Xiang, Shiming II-1  
 Xiao, Changyan I-418  
 Xiao, Fen II-159  
 Xiao, Feng II-419  
 Xiao, Liwu I-44

- Xiaogang, Zhang I-111  
 Xiaojun, Hu I-237  
 Xie, Jianhe I-363  
 Xie, Kaixuan I-325  
 Xie, Yaoqin II-392  
 Xiong, Ling I-21  
 Xu, Le II-97  
 Xu, Tingting II-22  
 Xu, Xiaopeng I-72
- Yan, Yunhui I-219  
 Yang, Bin II-107  
 Yang, Dawei I-355  
 Yang, Huanqing I-345  
 Yang, Jian I-140, II-341  
 Yang, Jie I-283  
 Yang, Jinfeng II-303  
 Yang, Jingwei II-585  
 Yang, Lu II-331  
 Yang, Meng II-311  
 Yang, Rui I-325  
 Yang, Shuang II-585  
 Yang, Shuo I-418  
 Yang, Wankou II-266, II-286  
 Yang, Wenming II-536  
 Yang, Yi II-545  
 Yang, Ze I-363  
 Yao, Cong I-391  
 Yao, Jiangang II-189  
 Yao, Lixiu I-283  
 Yibulayin, Tuergen II-491  
 Yin, Baolin I-455  
 Yin, Fei II-518  
 Ying, Gaoxuan I-445, II-199  
 Youguang, Zhang II-32  
 Yu, Haoming II-508, II-555  
 Yu, Hui II-128  
 Yu, Jun II-321  
 Yuan, Shuai II-255  
 Yuan, Xiuguang II-565  
 Yushan, Aliya I-411
- Zeng, Shirong II-228  
 Zhai, Shaozhuo I-293  
 Zhang, Baochang II-266  
 Zhang, Changshui I-11  
 Zhang, Danpu I-243  
 Zhang, David II-12
- Zhang, Defeng I-31  
 Zhang, Dewei II-565  
 Zhang, Gang I-426  
 Zhang, Guobao II-436  
 Zhang, Guoyun II-255  
 Zhang, Hongzhi II-12  
 Zhang, Jianhua I-345  
 Zhang, Jie II-138  
 Zhang, Lei II-555  
 Zhang, Mi II-189  
 Zhang, Rumin II-32  
 Zhang, Shaobo I-445  
 Zhang, Wentao II-392  
 Zhang, Xia II-89  
 Zhang, Xudong I-196  
 Zhang, Yanming I-1  
 Zhang, Yanxiang II-53  
 Zhang, Yongwang II-508  
 Zhang, Yuanping I-436  
 Zhang, Zhaoxiang I-314, II-117  
 Zhang, Zhenghui II-565  
 Zhao, Jiuliang I-219  
 Zhao, Qiyang I-455  
 Zhao, Wei II-508  
 Zhao, Weijian I-183  
 Zhao, Yang II-351  
 Zheng, Jun II-595  
 Zheng, Yali II-331  
 Zheng, Yuchen I-401  
 Zhong, Guoqiang I-401  
 Zhong, Qiang I-91  
 Zhou, Jiansong II-402  
 Zhou, Jin I-120  
 Zhou, Jiujiang II-149, II-179  
 Zhou, Xiaolong II-128  
 Zhou, Yan II-595  
 Zhou, Yongjin II-392  
 Zhou, Zhen II-276  
 Zhu, Dawei II-72  
 Zhu, Feiyue I-426  
 Zhu, Jiang I-54  
 Zhu, Liang I-91  
 Zhu, Qianlong I-44  
 Zhu, Songhao I-464  
 Zhu, Wenjie I-219  
 Zhu, Yuanping II-481  
 Zuo, Wangmeng II-12  
 Zuo, Xin II-286