

Changyin Sun Fang Fang
Zhi-Hua Zhou Wankou Yang
Zhi-Yong Liu (Eds.)

LNCS 8261

Intelligence Science and Big Data Engineering

4th International Conference, ISciDE 2013
Beijing, China, July/August 2013
Revised Selected Papers



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Changyin Sun Fang Fang Zhi-Hua Zhou
Wankou Yang Zhi-Yong Liu (Eds.)

Intelligence Science and Big Data Engineering

4th International Conference, IScIDE 2013
Beijing, China, July 31 – August 2, 2013
Revised Selected Papers

Volume Editors

Changyin Sun
University of Science and Technology, Beijing, China
E-mail: cys@ustb.edu.cn

Fang Fang
Peking University, Beijing, China
E-mail: ffang@pku.edu.cn

Zhi-Hua Zhou
Nanjing University, China
E-mail: zhouzh@nju.edu.cn

Wankou Yang
Southeast University, Nanjing, China
E-mail: wkyang@seu.edu.cn

Zhi-Yong Liu
Chinese Academy of Sciences, Beijing, China
E-mail: zhizhong.liu@ia.ac.cn

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-42056-6 e-ISBN 978-3-642-42057-3
DOI 10.1007/978-3-642-42057-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013952472

CR Subject Classification (1998): I.4.8, I.4, I.5.4, I.5, I.3, I.2.6, I.2.7, I.2.10, H.5.1, H.2.8, F.2.1-2

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

© Springer-Verlag Berlin Heidelberg 2013

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.

Typeetting: 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)

Preface

IScIDE 2013, the International Conference on Intelligence Science and Big Data Engineering, took place in Beijing, China, July 31 to August 2, 2013. As one of the annual events organized by the Chinese Golden Triangle ISIS (Information Science and Intelligence Science) Forum, this meeting was scheduled as the fourth of a series of annual meetings promoting the academic exchange of research on various areas of intelligence science and big data engineering in China and abroad. After three years of successful workshops, featuring over 300 submissions per meeting, an approximately 30% acceptance rate, and an oral presentation rate of about 5%, this meeting was formally changed to a conference. In response to the call for papers, a total of 390 papers were submitted from 13 countries and regions, of which 111 were accepted (giving an acceptance rate of about 28.5%), including 17 oral papers and 30 spotlight papers. We would like to thank all the reviewers for spending their precious time on reviewing papers and for providing valuable comments that aided significantly in the paper selection process.

We would like to express special thanks to the Conference General Chair, Lei Xu, for his leadership, advice, and help on crucial matters concerning the conference. We would like to thank all Steering Committee members, Program Committee members, Invited Speakers' Committee members, Organizing Committee members, and Publication Committee members for their hard work. We would like to thank Andrew Chi-Chih Yao and Lei Guo for delivering the keynote speeches, and Alireza Bab-Hadiashar, Joao Gama, Barry O'sullivan, Luc De Raedt, Rene Vidal, Zeng-Guang Hou, Shao Li, Yuanqing Li, Hanzi Wang, and Xiaokang Yang for delivering the invited talks and sharing their insightful views on ISIS research issues. Finally, we would like to thank all the authors of the submitted papers, whether accepted or not, for their contribution to the high quality of this conference. We count on your continued support of the ISIS community in the future.

August 2013

Fang Fang
Changyin Sun
Zhi-Hua Zhou
Wankou Yang
Zhiyong Liu

Organization

Steering Committee Chair

Lei Xu Chinese University of Hong Kong, Hong Kong

Steering Committee Members

Huafu Chen	University of Electronic Science and Technology of China
Fang Fang	Peking University
Xinbo Gao	Xidian University
Xiaofei He	Zhejiang University
Dewen Hu	National University of Defence Technology
James Kwok	Hong Kong University of Science and Technology
Xiang Li	Fudan University
Bao-liang Lu	Shanghai Jiao Tong University
Changyin Sun	University of Science and Technology Beijing
Xianglong Tang	Harbin Institute of Technology
Vincent Tseng	National Cheng Kung University Taiwan
Xihong Wu	Peking University
Jian Yang	Nanjing University of Science and Technology
Changshui Zhang	Tsinghua University
Zhi-Hua Zhou	Nanjing University

Steering Committee Secretary-General

Yanning Zhang Northwestern Polytechnic University

General Chairs

Lei Xu Chinese University of Hong Kong
Hong Mei Peking University

Program Committee Chairs

Fang Fang Peking University
Changyin Sun University of Science and Technology Beijing
Zhi-Hua Zhou Nanjing University

VIII Organization

Organizing Committee Chairs

Yi-Xin Yin	University of Science and Technology Beijing
Hui Zeng	University of Science and Technology Beijing
Xihong Wu	Peking University

Post-conference Workshop Chair

James Kwok	Hong Kong University of Science and Technology
------------	--

Publicity Chair

Zhouchen Lin	Peking University
--------------	-------------------

Local Arrangements Chair

Wankou Yang	Southeast University
-------------	----------------------

Registration Chairs

Chaoxu Mu	Southeast University
Yang Yi	Beihang University

Program Committee Members

Shu-Heng Chen	National Chengchi University, Taiwan
Seungjin Choi	Pohang University of Science and Technology South Korea
Jufu Feng	Peking University, Beijing
Kenji Fukumizu	The Institute of Statistical Mathematics, Japan
Xinbo Gao	Xidian University, Xi'an
Xiaofei He	Zhejiang University, Hangzhou
Kalviaainen Heikki	Lappeenranta University of Technology, Finland
Akira Hirose	The University of Tokyo, Japan
Tu Bao Ho	JAIST, Japan
Derek Hoiem	University of Illinois at Urbana-Champaign, 6USA
Dewen Hu	National University of Defence Technology, Changsha
Hiroyuki Iida	Japan Advanced Institute of Science and Technology, Japan
Ikeda Kazushi	Nara Advanced Institute of Science and Technology, Japan

James Kwok	Hong Kong University of Science and Technology, Hong Kong
Andrey S. Krylov	Lomonosov Moscow State University, Russia
Jian-huang Lai	Sun Yat-sen University
John Langford	Microsoft Research
Patrick Le Callet	Polytech Nantes/Université de Nantes, France
Minho Lee	Kyungpook National University, Korea
Xuelong Li	Xi'an Optics and Fine Mechanics, Chinese Academy of Sciences, Xi'an
Zhouchen Lin	Peking University, Beijing
Cheng Yuan Liou	National Taiwan University, Taiwan
Baoliang Lu	Shanghai Jiao Tong University, Shanghai
Seiichi Ozawa	Kobe University, Japan
Karl Ricanek	UNC Wilmington, USA
Cartic Ramakrishnan	USC/Information Science Institute, USA
Hichem Sahli	Vrije Universiteit Brussel, Belgium
Shiguang Shan	Institute of Computing Technology Chinese Academy of Sciences, Beijing
Hava Siegelmann	University of Massachusetts, USA
Fiori Simone	Universit Politecnica delle Marche, Italy
Vikas Sindhwani	IBM T.J. Watson Research Center, USA
Cox Stephen	University of East Anglia, Norwich, England
Dacheng Tao	University of Technology, Australia
Xianglong Tang	Har'bin Institute of Technology, Har'bin
Vincent Tseng	Cheng Kung University, Taiwan
Xihong Wu	Peking University, Beijing
Jian Yang	Nanjing University of Science and Technology, Nanjing
Qiang Yang	Hong Kong University of Science and Technology, Hong Kong
Changshui Zhang	Tsinghua University, Beijing
Kun Zhou	Zhejiang University, Hangzhou

Table of Contents

Advanced Variable Window Stereo Matching Algorithm	1
<i>Longyuan Guo, Changyin Sun, Guoyun Zhang, and Jianhui Wu</i>	
Unfocused Blur Assessment of SAR Images	6
<i>Han Zhang, Weidong Yan, Hui Bian, Weiping Ni, Junzheng Wu, Sha Li, Xinlu Ma, and Ying Lu</i>	
Syntactic Sensitive Complexity for Symbol-Free Sequence	14
<i>Cheng-Yuan Liou, Daw-Ran Liou, Alex A. Simak, and Bo-Shiang Huang</i>	
Compression in Molecular Simulation Datasets	22
<i>Anand Kumar, Xingquan Zhu, Yi-Cheng Tu, and Sagar Pandit</i>	
Online-Learning Structural Appearance Model for Robust Visual Tracking	30
<i>Min Yang, Mingtao Pei, Yuwei Wu, Bo Ma, and Yunde Jia</i>	
Polygon-Location Method Based on Uyghur Text Regional Rules	40
<i>Halidan Abudureyimu, Renren Deng, Kuerban Maitimusha, and Nana Yang</i>	
Study on the Electromagnetic Performance of Hydroelectric Generator Based on Intelligent Control	47
<i>Xin Shi, Jianhua Lin, Yuxiang Wang, and Heping Liu</i>	
Text-Independent Phoneme Segmentation via Learning Critical Acoustic Change Points	54
<i>Peng Teng, Xiabi Liu, and Yunde Jia</i>	
Robustness Analysis of Z-type ZLE Solving	62
<i>Weibing Li, Wenchao Lao, Xiaotian Yu, Zenghai Chen, and Yunong Zhang</i>	
Orthogonal Waveform Design Based on the Modified Chaos Genetic Algorithm for MIMO Radar	70
<i>Xin Fu, Xian Zhong Chen, Qing Wen Hou, Zhengpeng Wang, and Yixin Yin</i>	
Automatic Object Tracking in Aerial Videos via Spatial-temporal Feature Clustering	78
<i>Xiaomin Tong, Yanning Zhang, Tao Yang, and Wenguang Ma</i>	

A Neural Network for Parameter Estimation of the Exponentially Damped Sinusoids	86
<i>Xiuchun Xiao, Jian-Huang Lai, and Chang-Dong Wang</i>	
Local-Global Joint Decision Based Clustering for Airport Recognition	94
<i>Bingxin Qu, Yanning Zhang, and Tao Yang</i>	
Ultra-Wideband Interference Suppression in Time Reversal Transmitted-Reference UWB System	103
<i>Lan Zhang, Fang-Chao Zhang, and Bing Wang</i>	
Exploiting the Wisdom of Crowd: A Multi-granularity Approach to Clustering Ensemble	112
<i>Dong Huang, Jian-Huang Lai, and Chang-Dong Wang</i>	
Spatio-temporal Features for Efficient Video Copy Detection	120
<i>Ruijuan Hu, Bing Li, Weiming Hu, and Jinfeng Yang</i>	
Improve Scene Classification by Using Feature and Kernel Combination	128
<i>Lin Yuan, Fanglin Chen, Li Zhou, and Dewen Hu</i>	
Horror Text Recognition Based on Generalized Expectation Criteria	136
<i>Guoqi Liu, Bing Li, Weiming Hu, and Jinfeng Yang</i>	
A Representative-Sequence Based Near-Duplicate Video Detection Method	143
<i>Chen Shi, Li Zhuo, Yingdi Zhao, and Yuanfan Peng</i>	
Non-negative Sparse Coding Using Independent Multi-Codebooks for Near-Duplicate Image Detection	152
<i>Shan Zhou, Jun Li, Junliang Xing, Weiming Hu, and Jinfeng Yang</i>	
Machine Vision Based Automatic Micro-parts Detection System	160
<i>Xianchuan Yu, Guanyin Gao, Wu He, and Jindong Xu</i>	
Segment Based Depth Extraction Approach for Monocular Image with Linear Perspective	168
<i>Yiming Mo, Tianliang Liu, Xiuchang Zhu, Xiubin Dai, and Jiebo Luo</i>	
A Fuzzy Mix-Prototype Clustering Algorithm for Leukemia Data Analysis	176
<i>Jin Liu, Qianping Wang, Zhizhen Liang, and Wei Chen</i>	
Blind Image Quality Assessment with Semi-supervised Learning and Fuzzy Logic	184
<i>Ning Mei, Fei Gao, Wen Lu, and Xinbo Gao</i>	

Cross-View Action Recognition via Bilingual Bag of Dynamical Systems	192
<i>Changhong Chen, Shunqing Yang, and Zongliang Gan</i>	
Robust Image Set Classification Using Partial Least Squares	200
<i>Hui Jin and Ruiping Wang</i>	
Research on Quality Improvement of Polarization Imaging in Foggy Conditions	208
<i>Congli Li, Wenjun Lu, Song Xue, and Yongchang Shi</i>	
Human Interaction Recognition by Spatial Structure Models	216
<i>Jianzhai Wu, Fanglin Chen, and Dewen Hu</i>	
Robust Principal Component Analysis for Recognition	223
<i>Yu Chen and Jian Yang</i>	
Time-Varying Distributed Resource Allocation Based on Thermal Minority Game	230
<i>Jin Liu, Qianping Wang, Zhizhen Liang, and Wei Chen</i>	
Salient Object Detection via Fast Iterative Truncated Nuclear Norm Recovery	238
<i>Chuhang Zou, Yao Hu, Deng Cai, and Xiaofei He</i>	
An Efficient Resource Allocation Method for Multimedia Cloud Computing	246
<i>Yirui Li, Li Zhuo, and Haojie Shen</i>	
Research and Application of Corrosion Prediction Based on GRA-SVR	255
<i>Dongmei Fu, Jinlong Xiang, and Xiaogang Li</i>	
Research on a Super-Sparse Data Generation Model for Temperature Data Map	263
<i>Dongmei Fu, Wuchen Li, Xingen Li, and Xiaoming Wang</i>	
Strip Flatness and Gauge Multivariable Control at Cold Tandem Mill Based on Fuzzy RBF Neural Network	271
<i>Li Wang</i>	
Medical Image Segmentation Based on FCM and Wavelets	279
<i>Zhihong Shi, Yi Liu, and Qian Li</i>	
A Derivative Augmented Lagrangian Method for Fast Total Variation Based Image Restoration	287
<i>Dongwei Ren, Wangmeng Zuo, Hongzhi Zhang, and David Zhang</i>	

A Novel Multi-class Brain-Computer Interface (BCI) Paradigm Based on Motor Imagery Sequential Coding (MISC) Protocol.....	295
<i>Jun Jiang, Erwei Yin, Yang Yu, Jingsheng Tang, Zongtan Zhou, and Dewen Hu</i>	
Image Segmentation Based on NSCT and BF-PSO Algorithm	303
<i>Le Wang and Zhenbing Zhao</i>	
Non-linear Feature Fusion Based on Polynomial Correlation Filter for Face Recognition	312
<i>Dong Yan, Yuanyuan Shen, Yan Yan, and Hanzi Wang</i>	
Robust Head Pose Estimation with a New Principal Optimal Tradeoff Filter.....	320
<i>Dong Yan, Yan Yan, and Hanzi Wang</i>	
Automated Tongue Segmentation Based on 2D Gabor Filters and Fast Marching	328
<i>Zhenchao Cui, Wangmeng Zuo, Hongzhi Zhang, and David Zhang</i>	
Hyperspectral Medical Images Unmixing for Cancer Screening Based on Rotational Independent Component Analysis	336
<i>Bo Du, Nan Wang, Liangpei Zhang, and Dacheng Tao</i>	
GPCA on Gabor Tensor for Face Recognition	344
<i>Lian Zhu, Rong Huang, Xinfu Ye, Wankou Yang, and Sun Changyin</i>	
Nonnegative Discriminative Manifold Learning for Hyperspectral Data Dimension Reduction	351
<i>Lefei Zhang, Liangpei Zhang, Dacheng Tao, Xin Huang, and Bo Du</i>	
Spectral Unmixing for Hyperspectral Image Classification with an Adaptive Endmember Selection.....	359
<i>Qingjie Meng, Yanning Zhang, Wei Wei, and Lei Zhang</i>	
A Subarea-Location Joint Spelling Paradigm for the BCI Control	368
<i>Erwei Yin, Jun Jiang, Yang Yu, Jingsheng Tang, Zongtan Zhou, and Dewen Hu</i>	
A Robust Real-Time Tracking Method of Fast Video Object Based on Gaussian Kernel and Random Projection.....	376
<i>Yajuan Feng, Lina Wang, and Shiyin Qin</i>	
Harmonious Competition Learning for Gaussian Mixtures	385
<i>GuoJun Liu and XiangLong Tang</i>	
A Hierarchical Path Planning Approach Based on Reinforcement Learning for Mobile Robots	393
<i>Qi Guo, Lei Zuo, Rui Zheng, and Xin Xu</i>	

Color Image Segmentation Based-on SVM Using Mixed Features and Combined Kernel	401
<i>Lei Li, Dong yan Shi, and Jun Xu</i>	
Co-expressing Patterns of Schizophrenia Candidate Genes in Brain Regions	410
<i>Xinguo Lu, Bingtao Feng, Yong Deng, and Dewen Hu</i>	
Boosting Deformable Part Model by Sample Sharing and Outlier Ablation	418
<i>Feng Liu, Yongzhen Huang, Liang Wang, and Wankou Yang</i>	
Multi-cue Visual Tracking Based on Sparse Representation	427
<i>Xiping Duan, Jiafeng Liu, and XiangLong Tang</i>	
LMDA: Local Maximum Discrimination Analysis	435
<i>Jun Gao</i>	
Visual Saliency Detection via Homology Distribution and Color Contrast	441
<i>Zhihui Chen, Yan Yan, and Hanzi Wang</i>	
Multi-Modal Multiple-Instance Learning and Attribute Discovery with the Application to the Web Violent Video Detection	449
<i>Shuai Hao, Ou Wu, Weiming Hu, and Jinfeng Yang</i>	
Design and Implementation of a Bimodal Face Recognition System	457
<i>Yong Xu, Jian Yang, Jiajie Xu, Qi Zhu, and Zizhu Fan</i>	
The Translation-Invariant Metric and Its Application	465
<i>Bing Sun, Jufu Feng, and Guoping Wang</i>	
A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks Based Large Vocabulary Chinese Speech Recognition	473
<i>Xiangang Li, Yuning Yang, and Xihong Wu</i>	
Multi-level Linguistic Knowledge Based Chinese Grapheme-to-Phoneme Conversion	481
<i>Yi Liu, Xiaojun Chen, Caixia Gong, and Xihong Wu</i>	
Blind Quality Assessment on Binary Seal Images	489
<i>Chengyun Wang, Zhanlong Hao, and Youbin Chen</i>	
An Improved 3D Edge Surface Tracking Algorithm Based on 3D Fractional-Order Differentiation within Confocal Microscopy Images	497
<i>Yu Ma, Yanning Zhang, and Lisheng Wang</i>	
A Multiobjective Fuzzy Clustering Algorithm Based on Robust Local Spatial Information for Image Segmentation	505
<i>Feng Zhao, Hanqiang Liu, and Jiulun Fan</i>	

Image Super-Resolution Based on Data-Driven Gaussian Process Regression	513
<i>Yan-Yun Qu, Meng-Jie Liao, Yan-Wen Zhou, Tian-Zhu Fang, Li Lin, and Hai-Ying Zhang</i>	
Face Recognition Based on Non-Subsampled Contourlet Transform and Multi-order Fusion Binary Patterns	521
<i>Yao Deng, Weifeng Li, Zhenhua Guo, and Youbin Chen</i>	
Texture-Aware Fast Global Level Set Evolution	529
<i>Souleymane Balla-Arabé, Xinbo Gao, and Lai Xu</i>	
A Novel Metric for Image Denoising Algorithms	538
<i>Yingtao Zhang, H.D. Cheng, Jianhua Huang, and XiangLong Tang</i>	
Adaptive Weight Optimization for Classification of Imbalanced Data ...	546
<i>Wenhai Huang, Guojie Song, Man Li, Weisong Hu, and Kunqing Xie</i>	
Feature Selection via Sparse Regression for Classification of Functional Brain Networks	554
<i>Yilun Wang, Guorong Wu, Zhiliang Long, Jingwei Sheng, Jiang Zhang, and Huafu Chen</i>	
Efficient Euclidean Local-Structural Based Sparse Coding for Robust Visual Tracking	561
<i>Ji Zhang, Hong-Yuan Wang, and Fu-Hua Chen</i>	
An Efficient Semi-supervised Hashing Method Based on Graph Transduction	570
<i>Xiumei Wang, Xianjun Gao, Jie Li, and Ying Wang</i>	
Fuzzy-PI Switch Control in Intermediate Frequency Heating Process of 3PE-Coating	579
<i>Xu Yang, Bo Wen, Chao-nan Tong, Yu-zhi Feng, and Yan Zeng</i>	
An Improved Method in Change Detection of Multitemporal Remote Sensing Image	587
<i>Fangshun Liao, Sufen Yu, Ying Li, and Yanning Zhang</i>	
A Probability-Based Object Tracking Method	595
<i>Xu Song, Guoqiang Li, Ying Li, and Yanning Zhang</i>	
Pedestrian Detection Based on Incremental Learning	603
<i>Yu Xia, Yongzhen Huang, Liang Wang, and Xin Geng</i>	
Vanishing Point Detection Based on Infrared Road Images for Night Vision Navigation	611
<i>Huan Wang, Feifei Li, and Mingwu Ren</i>	

Neuro-control to Energy Minimization for a Class of Chaotic Systems Based on ADP Algorithm	618
<i>Ruizhuo Song, Wendong Xiao, and Qinglai Wei</i>	
Low SNR FMCW Signal Processing with Prior Information	626
<i>Qing Wen Hou, Zhi Wei Xu, Zhen Long Bai, Xian Zhong Chen, and Jing Ni Wang</i>	
Sparsity Preserving Score for Joint Feature Selection	635
<i>Hui Yan</i>	
Adaptive Backstepping Controller Design for Reentry Attitude of Near Space Hypersonic Vehicle	642
<i>Jingmei Zhang, Changyin Sun, Ruimin Zhang, Chengshan Qian, and Lei Xue</i>	
High Performance Super-Resolution Reconstruction of Multiple Images Based on Fast Registration and Edge Enhancement	649
<i>Ming Liu, Jianyu Huang, Ming Gao, and Shiyin Qin</i>	
Foreground Detection via Motion Field Based MRF-MAP	658
<i>Limin Zhu and Yue Zhou</i>	
The Speaker Recognition of Noisy Short Utterance	666
<i>Ying Chen and Zhen-Min Tang</i>	
Hyper-graph Matching with Bundled Feature	672
<i>Deyuan Li and Yue Zhou</i>	
Methods for Photomosaic Generation Based on Different Image Similarity and Division Strategies	682
<i>Daqing Chang, Changshui Zhang, and Shifeng Weng</i>	
Robust Continuous Terminal Sliding Mode Control Design for a Near-Space Hypersonic Vehicle	691
<i>Ruimin Zhang, Changyin Sun, Jingmei Zhang, and Chengshan Qian</i>	
A Level Set with Shape Priors Using Moment-Based Alignment and Locality Preserving Projections	697
<i>Bin Wang, Xinbo Gao, Jie Li, Xuelong Li, and Dacheng Tao</i>	
An Improved Texture Feature Extraction Method for Tyre Tread Patterns	705
<i>Ying Liu, Zong Li, and Zi-Ming Gao</i>	
Multi Gesture Recognition: A Tracking Learning Detection Approach . . .	714
<i>Meng-Yuan Shi and De-Chuan Zhan</i>	
Kernel-Based Representation Policy Iteration with Applications to Optimal Path Tracking of Wheeled Mobile Robots	722
<i>Zhenhua Huang, Xin Xu, Lei Ye, and Lei Zuo</i>	

XVIII Table of Contents

A Combined MRSMC/MBC Altitude Controller for a Quad-rotor UAV	731
<i>Wei Wang, Hao Ma, and Changyin Sun</i>	
PerGrab: Adapting Grabbing Gesture Recognition for Personalized Non-contact HCI	740
<i>Tao Li and Ming Li</i>	
An Automatic MSRM Method with a Feedback Based on Shape Information for Auroral Oval Segmentation	748
<i>Hui Liu, Xinbo Gao, Bing Han, and Xi Yang</i>	
An Improved ELM Algorithm Based on EM-ELM and Ridge Regression	756
<i>Haigang Zhang, Sen Zhang, and Yixin Yin</i>	
An Attitude Determination System of Quad-rotor Aircraft Based on Extended Kalman Filter and Data Fusion Technique	764
<i>Xinfu Ye, Lian Zhu, Sun Changyin, and Wei Wang</i>	
Connectivity of Clustered and Multi-type User CR Network: A Percolation Based Approach	771
<i>Jingyuan Guo, Tao Yang, Hui Feng, and Bo Hu</i>	
Patch-Based Tracking and Detecting for Visual Tracking	779
<i>Qianwen Li and Yue Zhou</i>	
Adaptive Regularization Parameters and Norm Selection for Sparse Gradient Based Image Restoration	789
<i>Xinqian Lin, Hongzhi Zhang, Hong Deng, and Wangmeng Zuo</i>	
Key-Frame Selection Strategy Based on Edge Points Classification in 2D-to-3D Conversion	797
<i>Jiangchuan Xie, Jiande Sun, Ju Liu, and Qiaoli Hu</i>	
Camera Localization and Pose Estimation Using an RGBD Sensor	805
<i>Hao Chen and Yan Yuan</i>	
Sparse Brain Anatomical Network Based Classification of Schizophrenia Patients and Healthy Controls	813
<i>Junjie Zheng, Yilun Wang, Heng Chen, and Huafu Chen</i>	
Sparse Learning for Face Recognition with Social Context	820
<i>Jie Gui, Jian-Xun Mi, Ying-Ke Lei, and Hong-Qiang Wang</i>	
Finger Vein Recognition Based on Gabor Filter	827
<i>Hong Zhang, Zhi Liu, Qijun Zhao, Congcong Zhang, and Dandan Fan</i>	
Construction and Simulation Analysis for Stability Speed Parameter of Instantaneous Availability for One-Unit Repairable Systems	835
<i>Yi Yang, Lichao Wang, and Rui Kang</i>	

Compressed Sensing Ensemble Classifier for Human Detection	844
<i>Baochang Zhang, Juan Liu, Yongsheng Gao, and Jianzhuang Liu</i>	
A Prediction Reference Structure Based Hierarchical Perceptual Encryption Algorithm for H.264 Bitstream	852
<i>Haojie Shen, Li Zhuo, and Yirui Li</i>	
A New Image Denoising and Enhancement Method Combining the Nonsubsampled Contourlet Transform and Improved Total Variation	860
<i>Ying Li, Yu Jia, and Yanning Zhang</i>	
An Improved Particle Swarm Optimization for Complex Optimization Problems	868
<i>Kezong Tang, Binxiang Liu, and Jia Zhao</i>	
An Improved Method for Oriented Chamfer Matching	875
<i>Jian Dong, Changyin Sun, and Wankou Yang</i>	
Computer Vision Based Pose Bias Detection of Shield Tunneling Machine	880
<i>Jiannan Chi, Lei Liu, Jiwei Liu, Changyin Sun, and Weiping Zhang</i>	
Integrative Hypothesis Test and A5 Formulation: Sample Pairing Delta, Case Control Study, and Boundary Based Statistics	887
<i>Lei Xu</i>	
Erratum	
Camera Localization and Pose Estimation Using an RGBD Sensor	E1
<i>Hao Chen, Yan Yuan, John McDonald, and Thomas Whelan</i>	
Author Index	903

Advanced Variable Window Stereo Matching Algorithm

Longyuan Guo^{1,2}, Changyin Sun², Guoyun Zhang¹, and Jianhui Wu¹

¹ Hunan Institute of Science & Technology,

Key Laboratory of Optimization and Control for Complex Systems,

College of Hunan Province, Yueyang 414006, China

guolongyuan@hotmail.com

² School of Automation, Southeast University, Nanjing 210096, China

cysun@seu.edu.cn

Abstract. Variable window stereo matching methods overcome the disadvantages of fixed window methods. Using this base idea, an advanced Variable window stereo matching method is proposed. The method takes a certain threshold to determine matching windows. In order to improve the accuracy of matching further, a cost function using the non-parametric and gray value is proposed. The experimental results show that this method can generates more accurate disparity map.

Keywords: Variable window, Area-based matching, Census.

1 Introduction

Binocular stereo matching is an important research direction in computer vision. Area-based matching methods, which can directly generate a dense disparity map, are applied widely. However, match window size and shape will impact on the matching result directly. Therefore, the variable and adaptive window matching methods were proposed [1-3]. Kanade and Okutomi [3] first proposed a method of adaptive window, which using an initial disparity estimation value and selecting a different window for iterative matching until the data convergence. The drawback of this method is the large amount of calculation and results were affected by the initial disparity. Veksler [4] adopts non-rectangular windows, with the lowest proportion of circulating algorithm, to optimize a large class of compact window to select the shape of the window. However, this algorithm is too complex and not suitable for real-time system.

Yoon and Kweon [5] proposed a local adaptive support weight matching method, the weight of each pixel is related to dissimilarity of the color and spatial distance to center pixel. This method is easy to produce the image noise.

Learning ZHANG's method [6], this article proposed a novel matching support window, which based on grayscale difference between the surrounding and the center pixel. At same time, a composite cost function value is combined to non-parametric measure and gray values. After occlusion detection, original matching disparity map will become more accurate.

2 Matching Window

Area-based matching method is based on a hypothesis, i.e. the disparity of a pixel is equal to adjoining region pixels'. So, a region center in the pixel can be used as a measure of finding matching points in another image.

It's difficulties to determine size of the relevant window. If matching window's size is too large, the discontinuities disparity will become blurring; If the size is too small, the multiple optimal matches will appear in smooth gray region. In this paper, we focus on the cross-based aggregation method proposed firstly by Zhang et al[6]. Cross-based aggregation proceeds by a two-step process. In the first step, an upright cross with four arms is constructed for each pixel. In the second step, the aggregated costs over all pixels are computed within two passes: the first pass sums up the matching costs horizontally and stores the intermediate results; the second pass then aggregates the intermediate results vertically to get the final costs.

This article inspired by it and proposes the method which comparing pixel gray value within a certain range and less than a certain threshold value of the pixel to be included in the matching window. The pixels in the matching window should satisfy as following:

$$|I(x, y) - I_0(x_0, y_0)| < \lambda \quad (1)$$

Where, $I_0(x_0, y_0)$ is the center pixel gray value. $I(x, y) \in N_{I_0}$, N_{I_0} is a set of matching window pixels gray value; λ is gray threshold.

With this method, the matching window can be determined quickly and improve the matching speed.

Meanwhile, in order to ensure the gray value near the center greater effect and avoid influence of distant pixels, each pixel gray values are multiplied by the weight value during window aggregation proceeding.

3 Cost Function

The SAD algorithm cost function is simply constituted by pixel gray value. While the gray value usually is affected by the noise and other factors. This article draw on Mei's[7] constructed cost function, and integrate gray value difference and Census Hamming distance as total cost function value.

Census is a non-parametric transform mathematical statistics method. The basic idea is that to compare the detection unit and reference unit and statistics to determine whether the signal is present [8]. Assuming $I(x, y)$ is gray value of the pixel (x, y) ; $N(x, y)$ denotes a pixel set in window centered on (x, y) . Then pixel (x, y) Census transform value in $N(x, y)$ domain is defined as:

$$R(x, y) = \underset{\xi, \eta \in N}{\text{BitString}} \delta(I(x, y), I(x + \xi, y + \eta)) \quad (2)$$

BitString connect one after the other

$$\delta(\alpha, \beta) = \begin{cases} 0 & \beta \leq \alpha \\ 1 & \beta > \alpha \end{cases}$$

Census value reflects the relative order of the local image gray value, rather than gray value of the pixel. Thereby the effect will reduce on the image by radiation distortion and noise.

However, Census values is ambiguous in repeat texture or structure region. Gray values can provide a more subtle information. Therefore, it is conducive for solving the ambiguous to combine the gray information. Cost function is defined as:

$$C(x, y) = 2 - \exp\left(-\frac{C_{SAD}}{k_{SAD}}\right) - \exp\left(-\frac{C_{census}}{k_{census}}\right) \quad (3)$$

Where, grayscale measure using the SAD algorithm,

$$C_{SAD}(x, y) = \sum_{\xi, \eta \in N} |I_1(x + \xi, y + \eta) - I_2(x + \xi + k, y + \eta)| \quad (4)$$

Census Hamming measure is,

$$C_{census} = \sum_{(\xi, \eta)} \delta_r(I(x_2, y_2), I(x_2 + \xi, y_2 + \eta) \oplus \delta_l(I(x_1, y_1), I(x_1 + \xi, y_1 + \eta)) \quad (5)$$

k_{SAD} and k_{census} , two constants, are used to keep the cost value greater than 0, and adjust two measures' ratio in the total measure. After obtaining the original disparity map with the above cost function, the paper using the left and right consistency check and the sequential to reduce mismatching further.

4 Experimental Result and Discussion

In order to test the effect of the method, the paper using the picture provided by Middlebury stereo matching evaluation system, and simulation in Matlab.

The experimental results are shown in Figure 1. Many false match can be seen in Tsukuba disparity generated by SAD algorithm, especially in occlude. For example, the mismatching areas, at the left edge of statues and the jar at the table, are obvious. While, the proposed algorithm eliminates these mismatching. Mismatching area of Aloe due to the occlusion in the leaves, the results of the proposed algorithm is also ideal.

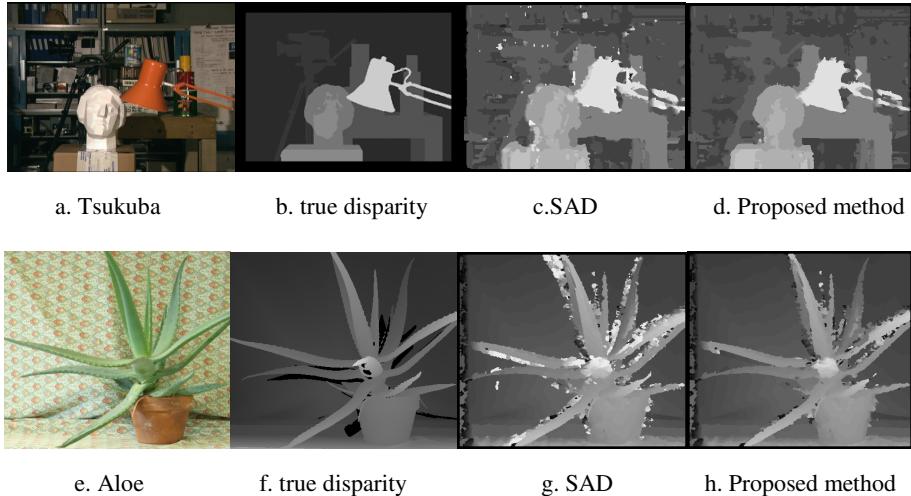


Fig. 1. Comparison of stereo matching result

5 Conclusion

Area-based variable window matching method overcomes the disadvantage of the fixed window matching method. With fixed window, if the correlation window is too large, the disparity of discontinuities become blurring; if the window is too small, there are multiple matches in a region with a uniform gray level. The paper adopted variable window base idea, and proposed a method which can generate matching window quickly. That is, the pixels in the window are bound to meet a certain threshold. At same time, the paper combined non-parametric and gray value as the cost function. The experimental results show that the present method can improves the matching correct rate and has good matching effect.

Acknowledgements. This research was supported by Projects of Hunan Province Science & Technology Department (2013GK3097). Youth project of National Natural Science Fund No.61201435, Education Department of Hunan Province science and technology project No.10A046 and Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province. Its contents are solely the responsibility of the authors and do not necessarily represent the official views.

References

1. Zhou, X., Wen, G., Wang, R.: Fast stereo matching using adaptive window. *Chinese Journal of Computers* 29(3), 473–479 (2006)
2. Weiji, N., Guili, X., Yupeng, T., Biao, W., Xin, C.: Fast stereo matching based on color segmentation and adaptive window. *Chinese Journal of Scientific Instrument* 32(1), 194–200 (2011)

3. Kanade, T., Okutomi, M.: A stereo matching algorithm with adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(9), 920–932 (1994)
4. Veksler, O.: Stereo matching by compact window via minimum ratio cycle. *International Journal of Computer Vision* 1, 556–561 (2002)
5. Yoon, K.J., Kweon, I.S.: Adaptive support weight approach for stereo correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(4), 650–656 (2005)
6. Zhang, K., Lu, J., Lafruit, G.: Cross-based local stereo matching using orthogonal integral images. *IEEE TCSVT* 19(7), 1073–1079 (2009)
7. Mei, X., Sun, X., Zhou, M., Jiao, S., Wang, H., Zhang, X.: On building an accurate stereo matching system on graphics hardware. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 467–474 (2011)
8. 段凤增编著,信号检测理论. 哈尔滨:哈尔滨工业大学出版社, 226–228 (2002)

Unfocused Blur Assessment of SAR Images

Han Zhang, Weidong Yan, Hui Bian, Weiping Ni, Junzheng Wu, Sha Li,
Xinlu Ma, and Ying Lu

Northwest Institute of Nuclear Technology, 710024 Xi'an, China
zhanghan9718@163.com

Abstract. Unfocused blur assessment is an important part of SAR image quality evaluation system. The mainlobe of unfocused point target is widened and the sidelobe is raised, leading to broadened edges in the azimuth direction. Based on this phenomenon, we proposed a new method to evaluate the extent of unfocused blur using truncated image spectrum and the average edge width in the salient area. A new metric UBE is defined and verified by experiments.

Keywords: SAR Image, unfocused blur, Radon transform, saliency.

1 Introduction

SAR images are subject to a wide variety of distortions during acquisition and processing, such as geometric distortion, radiometric error, ambiguity, sidelobes, clutter noise and unfocused blur [1]. Unfocused blur is caused by error analysis of Doppler frequency rate. The parameter dependency of Doppler frequency rate is a function of the velocity of radar platform. Therefore, the platform stability and consistent velocity of moving radar sensor is to be guaranteed for the fine image resolution. In practice, it may be difficult to maintain those requirement, thus an accurate estimation of the relative velocity is required to have minimum estimation error [2]. The Doppler frequency rate governs the phase of the azimuth matched filter. The presence of an error in Doppler frequency rate causes filter mismatch, which could bring about the following imaging distortion.

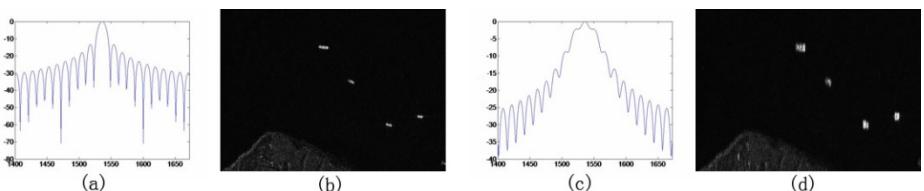


Fig. 1. Focused and unfocused SAR images, (a) azimuth slice of focused point target, (b) focused SAR image, (c) azimuth slice of unfocused point target, (d) unfocused SAR image

Fig.1 shows imaging results of the focused and unfocused SAR images. The unfocused one is caused by 6% estimation error of Doppler frequency rate. We can see

that the unfocused point target image has widened mainlobe and raised sidelobe. Edges of the blurred image are widened in the azimuth direction as that caused by motion blur. We identify unfocused blur extent here with a new method based on the motion blur estimation.

Studies on SAR image unfocused blur assessment have not been investigated widely. Y. Yitzhaky developed a method based on the concepts that the smoothness of the blurred image in the motion direction are greater than in other directions and that, correlation existing in this direction between the pixels forms the blur of the original unblurred objects [3]. Combined with the averaging operation, R. Zhang used the correlation method to measure the blur extent of SAR images [4]. The inspection of zero patterns of blurred images in the spectral domain has been proposed to find the blur direction and the blur extent of uniform velocity motion [5-7]. In this paper, characteristics of unfocused blur SAR image spectrum are analysed and the truncated spectrum is used to identify the unfocused direction, then the average edge width in the azimuth direction of the salient image area is used to estimate the extent of unfocused blur. A block diagram summarizing the new method is shown in Fig.2. The contributions of this paper are that we use the truncated spectrum, large window median filter and saliency to overcome the influence of speckles, and proposed an effective unfocused blur metric UBE.

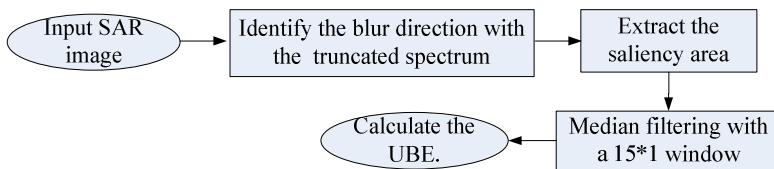


Fig. 2. Block diagram summarizing the proposed method

2 Unfocused Blur Direction Identification

The unfocused blur direction should be identified before the estimation of blur extent, and here, the Radon transform of truncated spectrum is adopted.

2.1 Spectrum of Unfocused SAR Image

We take the uniform velocity motion kernel to describe the unfocused blur of SAR image. The blur function is given as follows [7]:

$$g(x, y) = f(x, y) * h(x, y) + n(x, y) \quad (1)$$

where $g(x, y)$ is the blurred image, $f(x, y)$ is the focused image, $h(x, y)$ is the motion kernel, and $n(x, y)$ is the additive noise. $h(x, y)$ is given as:

$$h(x, y) = \begin{cases} \frac{1}{L} & \text{if } \sqrt{x^2 + y^2} \leq \frac{L}{2}, \frac{x}{y} = -\tan(\theta) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where L is the blur extent, θ is the blur direction.

In the frequency domain, Eq. (1) is given by:

$$G(u, v) = F(u, v)H(u, v) + N(u, v) \quad (3)$$

where

$$H(u, v) = \sin c(L(u + v \tan \theta) / 2) \quad (4)$$

as shown in Fig. 3.

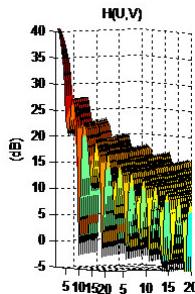


Fig. 3. Spectrum of uniform velocity motion kernel

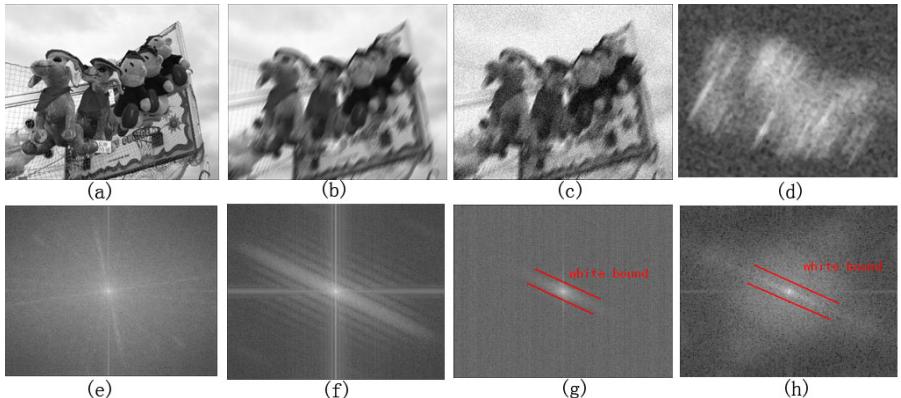


Fig. 4. Spectrum of blur images, (a) clear optical image, (b) motion blurred optical image, (c) motion blurred noisy optical image, (d) unfocused SAR image, (e) spectrum of (a), (f) spectrum of (b), (g) spectrum of (c), (h) spectrum of (d)

We can see that $H(u, v)$ is a 2-dimension SINC function in which parallel dark lines exist in the positions $u + v \tan \theta = 2\pi i / L$, $i = 1, 2, \dots, N$. These lines are vertical to the blur direction θ . Eq. (3) shows that the spectrum of unfocused blur image is got by multiplying the spectrum of the corresponding focused image by the blur kernel $H(u, v)$ and then adding to the noise spectrum $N(u, v)$. Ignoring the additive noise, the parallel dark lines should also be found on the spectrum of blurred image. Thus, the blur direction and blur extent can be calculated from the spectrum of blurred image using Eq. (4).

Fig. 4 (a) was obtained from the LIVE database [9] and Fig. 4 (b) is the uniform velocity motion blurred image with $\theta = 60^\circ, L = 20$. Fig. 4 (c) is the motion blurred noisy image with 10 dB Gaussian noise. Fig. 4 (d) is an unfocused SAR image caused by 6% percent Doppler frequency rate error. Fig. 4 (e)-(h) is the corresponding spectrum image. We can see that the spectrum of motion blurred noise free optical image consists of apparent parallel dark lines as mentioned above and we can identify the motion blur parameters from the spectrum, while the spectrum of noisy motion blurred optical image and the unfocused blurred SAR image does not have such parallel dark lines. But we can still find a white bound in the center of the spectrum image that is vertical to the blur direction, and can be extracted to identify the blur direction.

2.2 Radon Transform of Truncated Spectrum

The Radon transform is widely used to extract straight lines from noisy images [5-7]. The Radon transform of image $I(x, y)$ is defined by

$$R_I(\rho, \theta) = \int_{-\infty}^{\infty} G(\rho \cos \theta - s \sin \theta, \rho \sin \theta + s \cos \theta) ds \quad (5)$$

which integrates G over a line of distance ρ from the origin and at an angle θ to the y -axis. It is easy to see that any line in the image will be represented by a peak in the Radon transform whose location determines the parameters of the line in the original image.

Recall that given an unfocused blur SAR image I and its spectrum F , we have white bound in $\log F$ with slope vertical to the blur direction θ_0 as shown in Fig. 4. The image $R_I(\rho, \theta)$ resulted from a Radon transform on $\log F$ should have a peak located at $\theta = \pi/2 + \theta_0$.

The Radon transform of unfocused blur SAR image spectrum in Fig. 4(h) is shown in Fig. 5(b). Because of the interruption of noise, no obvious peak exists in the Radon transform. To overcome this problem, we proposed a new method to extract the unfocused blur direction using the truncated image spectrum. A threshold T is settled and pixels smaller than T are set to zero, then we get the truncated spectrum as shown in Fig. 5(c). The Radon transform of truncated spectrum is shown in Fig. 5(d). An apparent peak is shown at the position θ_0 which is the unfocused blur direction.

The key point of the truncated spectrum method is to settle the threshold T . A detailed description of this process is as follows:

Given a SAR image spectrum F , $G = \log F$.

- (a) Calculate the mean of G as a , the maximum pixel value as m , set initial value $T = a$, step length $step = (m - a) / 20$.
- (b) Set $T = T + step$. Calculate the Radon transform of the truncated spectrum, and search for the first peak P_1 and the second peak P_2 .
- (c) If $(P_1 - P_2) / P_1 < 0.3$, return to step (b); Otherwise, stop searching and set the threshold as the current T .

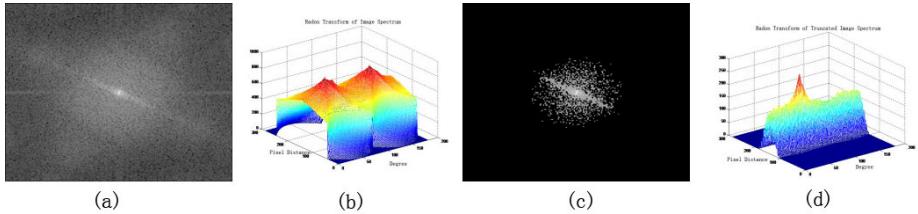


Fig. 5. Radon transform of truncated SAR image spectrum, (a) spectrum of blurred SAR image, (b) radon transform of (a), (c) truncated spectrum, (d) radon transform of (c)

3 Unfocused Blur Extent Identification

Our approach is relied on the observation that the unfocused blur SAR image has widened mainlobes and raised sidelobes, leading to broadened edges in the azimuth direction. The edge widened extent reflects the unfocused blur extent. However, the speckle noise would interrupt the process of edge locating and edge wide evaluation. We proposed to use the average edge wide in the salient area of SAR image after median filtered as a new measure of unfocused blur extent.

3.1 Salient Area of SAR Image

Gills defines saliency as the local signal complexity and uses the local signal entropy to measure it [8]. For SAR image, the salient area contains most of the brightest targets of big RCS, which will weaken the speckle influence on the edge locating and edge width evaluation. For a local region R in image I , calculate its gray histogram $P_R(d)$ where d is the gray value ranged in $[0, D]$. The local entropy H_R is defined as:

$$H_R = -\sum_{d=1}^D P_R(d) \log_2 P_R(d) \quad (6)$$

A detailed description of the saliency extraction process is as follows:

Given the original image $I(x, y)$ size of $[M, N]$, and the saliency image $S(x, y)$ of the same size.

- (a) Set the sliding window length $l = \min\{M / 8, N / 8, 64\}$;
- (b) For each pixel of I , calculate the local entropy $E(x, y)$, and get the entropy map E . Calculate the mean of E as E_{av} , set the threshold $T = E_{av}$;
- (c) Set $count = 0$, $S(x, y) = 0$;
- (d) For each pixel of E , if $E(x, y) > T$, take $I(x, y)$ as a salient point, set $S(x, y) = I(x, y)$, $count = count + 1$;
- (e) If $count / (M * N) > 0.2$, set $T = T + 0.1$, and return to step (c); otherwise stop the iterate.

Fig. 6(a) is an unfocused blur SAR image of Ottawa region, the saliency is shown in Fig. 6(b).

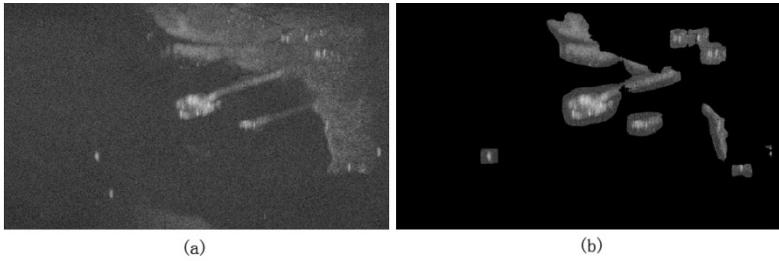


Fig. 6. Saliency extraction, (a) SAR image, (b) salient area of (a)

3.2 Median Filter of Large Window Size

Fig. 7(a) is the azimuth slice of Fig. 6(a). The image edges are interrupted by speckles, and the gray levels fluctuate very fast which make it impossible to locate the real image edges. Here we choose a one-dimension median filter of large window size for despeckling. The window size is set as 15·1 and the filtering direction is along the azimuth direction. Filtering result of Fig. 7(a) is shown in Fig. 7(b). The speckles are depressed effectively and the salient edges are saved.

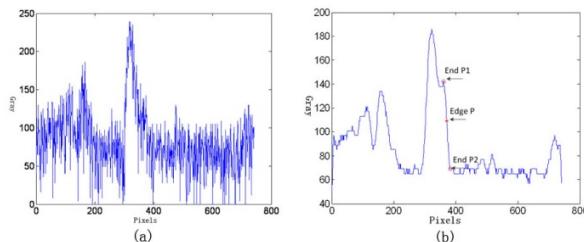


Fig. 7. Median filtering of large window size, (a) azimuth slice of SAR image, (b) azimuth slice after median filtering

3.3 Average Edge Width (AEW)

The Sobel operator is adopted to locate the edge points, such as point P in Fig. 7(b). The edge width of P is defined as the pixel length between the two ends P_1 and P_2 where the line P_1P_2 is monotonic. Given a SAR image $I(x, y)$, the edge points set is E of N elements in salient area. For each edge point p_i in E , the edge width in the azimuth direction is w_i . The average edge width AEW in salient area is defined as:

$$\text{AEW} = \frac{1}{N} \sum_{i=1}^N w_i \quad (7)$$

3.4 Unfocused Blur Extent (UBE)

We calculate the AEW of 100 SAR images of various kinds of landscape attained from different sensors. It is found that all the AEWs range between 20 and 30. We explain this phenomena with two reasons:

- (a) Because of the interruption of speckles, there would not be large area of the same or monotonic gray levels in SAR images. So the AEW would not be too large.
- (b) The interferential imaging method of SAR images and the median filtering process with a large window size of 15·1 make AEW no less than 20.

Based on the previous phenomena, we define a new unfocused blur extent measure UBE :

$$\text{UBE} = |\text{AEW} - 25| \quad (8)$$

and assess the unfocused blur extent by the following rules: if $\text{UBE} < 5$, identify the image as no blur; else if $5 \leq \text{UBE} \leq 10$, identify the image as slight blur; else if $\text{UBE} > 10$, identify the image as serious blur.

4 Results and Discussion

Two groups of SAR images are used to do the unfocused blur assessment experiments. Images of different unfocused blur extent caused by different Doppler frequency estimation error are shown in Fig. 8.

For the first group images of Fig. 8(a)-(c), when $e = 0\%$, $\text{UBE} = 1.6$, the image is determined to be no blur; when $e = 3\%$, $\text{UBE} = 6.7$, the image is determined to be slight blur; when $e = 6\%$, $\text{UBE} = 11.3$, the image is determined to be serious blur. For the second group images of Fig. 8(d)-(f), when $e = 0\%$, $\text{UBE} = 2.8$, the image is determined to be no blur; when $e = 3\%$, $\text{UBE} = 8.3$, the image is determined to be slight blur; when $e = 6\%$, $\text{UBE} = 12.2$, the image is determined to be serious blur. The determination is effective to describe the unfocused blur extent of SAR images.

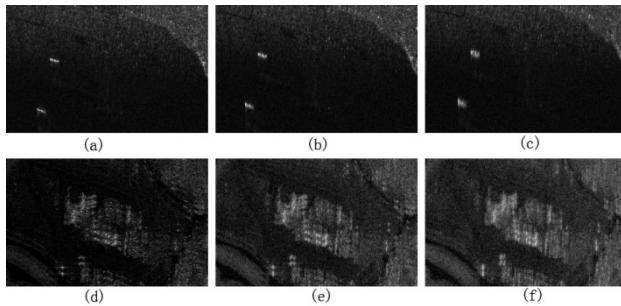


Fig. 8. UBE of different blur extent images, (a) $e = 0\%$, UBE = 1.6, (b) $e = 3\%$, UBE = 6.7, (c) $e = 6\%$, UBE = 11.3, (d) $e = 0\%$, UBE = 2.8, (e) $e = 3\%$, UBE = 8.3, (f) $e = 6\%$, UBE = 12.2

5 Conclusion

In this work, we proposed a new measure of SAR image unfocused blur extent (UBE). UBE is defined based on the observation that point targets of unfocused SAR image have widened mainlobes and raised sidelobes. As a result, the edge width of unfocused blur image is broadened. UBE is induced from the average edge width. In order to overcome the influence of speckles on direction determination and edge locating, the truncated image spectrum is used to identify the blur direction and the average edge width of the image salient area after one dimension median filtering is used to assess the blur extent. The method is verified by experiments.

References

1. Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images. SciTech Publishing, Inc., Raleigh (2004)
2. Jung, C.H., Choi, M.S., Kwag, Y.K.: Parameter Based SAR Simulator for Image Quality Evaluation. In: IEEE International Geoscience and Remote Symposium, pp. 1599–1602. IEEE Press, New York (2007)
3. Yitzhaky, Y., Kopeika, N.S.: Identification of Blur Parameters from Motion Blurred Images. In: Graphical Models and Image Processing, pp. 310–320 (1997)
4. Zhang, R., Yang, J.C., Zhang, Q., Liu, Z.K.: Motion Blur Extent Evaluation of SAR images. Acta Electronica Sinica 35(10), 2019–2022 (2007)
5. Chang, M.M., Tekalp, A.M., Erdem, A.T.: Blur Identification Using the Bispectrum. IEEE Transactions on Acoust, Speech, Signal Processing 39, 2323–2325 (1991)
6. Moghaddam, M.E., Jamzad, M.: Motion Blur Identification in Noisy Images using Mathematical Models and Statistical Measures. Pattern Recognition Society 40, 1946–1957 (2007)
7. Ji, H., Liu, C.Q.: Motion Blur Identification from Image Gradients. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, New York (2008)
8. Gilles, S.: Robust Description and Matching of Images. University of Oxford Press (1998)
9. Sheikh, H.R., Bovik, A.C., Cormack, L., Wang, Z.: LIVE Image Quality Assessment Database (2003), <http://live.ece.utexas.edu/research/quality>

Syntactic Sensitive Complexity for Symbol-Free Sequence

Cheng-Yuan Liou^{1,*,**}, Daw-Ran Liou¹, Alex A. Simak^{1,2},
and Bo-Shiang Huang¹

¹ Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan, Republic of China
cyliou@csie.ntu.edu.tw

² Institute of Statistical Science, Academia Sinica, Taiwan, Republic of China

Abstract. This work uses L-system to model the text sequence. The sequence complexity is obtained by calculating the complexity of its modeling system. It can sense certain quasi-regular structures and serves as a measure of regularity of the sequence. The outliers and statistics of the complexity values can be applied to the anomaly detection of symbol sequences.

Keywords: text complexity, quasi-regular structure, rewriting rule, L-system, measure of regularity.

1 Introduction

The entropy (or complexity) of the context-free language has been developed with varying degrees of success [1][2]. This entropy is applied to the estimation of loading capacity in communication transmission. Given a text sequence, it is hopeless to find a language that matches this sequence and compute its entropy. Since the Lindenmayer system [3], or L-system, can be used to model the sequence, the complexity of the sequence can be obtained indirectly by calculating the complexity of its modeling L-system. Such modeling complexity can be applied to text mining techniques, such as picking certain interesting sections in the sequence. This kind application is very different from the estimation of communication capacity. This paper shows how to use L-system to model the symbol sequence and derive its complexity.

Given a text, we first encode it into a binary string. Then, use L-system to model the tree structure of this string and get its modeling complexity. We will introduce how to use L-system to model the string in this section. The complexity for the text sequence is included in the next section.

1.1 Transforming Binary String into Rewriting Rules

L-system, is a parallel rewriting system which was introduced by the biologist Aristid Lindenmayer in 1968. The major operation of L-system is rewriting. A

* Corresponding author.

** Supported by NSC 100-2221-E-002-234-MY3 and NTU-EECS-102R3401-1.

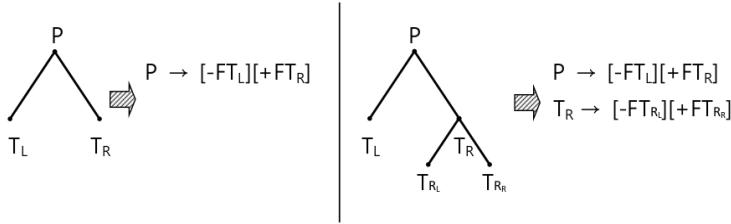


Fig. 1. Rewriting rules for the two bracketed strings

set of rewriting rules, or productions, are operated to define a complex object by successively replacing parts of a simple initial object. The operations for a hierarchical tree can be represented by a set of rewriting rules. These rules can be further transformed into a bracketed string. A binary tree can be represented by a bracketed string. Also, a tree can be restored from a string. This bracketed string contains five symbols, F , $+$, $-$, $[$, and $]$. These symbols are defined in below.

- F denotes the current location of a tree node. It can be replaced by any word or be omitted.
- $+$ denotes the following string that represents the right subtree.
- $-$ denotes the following string that represents the left subtree.
- $[$ is pairing with $]$. “[...]" denotes a subtree where “...” indicates the whole bracketed string of its subtree.

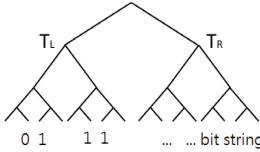
Given a binary tree, the direct way to represent it with L-system is to construct rewriting rules that replace a tree with two smaller subtrees. An example is shown in Fig. 1. The left tree in Fig. 1 is expressed as $P \rightarrow [-FT_L][+FT_R]$. The right tree is expressed as $P \rightarrow [-FT_L][+FT_R]$ and $T_R \rightarrow [-FT_{R_L}][+FT_{R_R}]$. We see that the two rules, $P \rightarrow [-FT_L][+FT_R]$ and $T_R \rightarrow [-FT_{R_L}][+FT_{R_R}]$, are similar. The binary tree example in Fig. 2 can be transformed into the set of rewriting rules in Table 1.

Fig.2 shows an example with four fixed tree elements. In this figure, every two leaves are combined into a small tree, and two small trees are combined into a bigger tree successively. In this way, we can get a whole binary tree for a long binary string. Each node of the tree represents all its descendant string sections. The binary tree example in Fig. 2 can be transformed into the set of rewriting rules in Table 1.

1.2 Classifying Rewriting Rules into Different Sets

In order to collect similar trees for classification, two similarity definitions are used in classifying the rewriting rules.

Definition 1. *Homomorphism in rewriting rules.* We say rewriting rule R_1 and rewriting rule R_2 are homomorphic to each other if and only if they have the same structure.

**Fig. 2.** Binary string represented by binary tree**Table 1.** Rewriting Rules for the binary tree in Fig.2

$P \rightarrow [-FT_L][+FT_R]$	$T_R \rightarrow [-FT_{R_L}][+FT_{R_R}]$
$T_L \rightarrow [-FT_{L_L}][+FT_{L_R}]$	$T_{R_L} \rightarrow [-FT_{R_{L_L}}][+FT_{R_{L_R}}]$
$T_{L_L} \rightarrow [-FT_{L_{L_L}}][+FT_{L_{L_R}}]$	$T_{R_{L_L}} \rightarrow [-F][+F]$
$T_{L_{L_L}} \rightarrow [-F][+F]$	$T_{R_{L_R}} \rightarrow [-F][+F]$
$T_{L_{L_R}} \rightarrow [-F][+F]$	$T_{R_R} \rightarrow [-FT_{R_{R_L}}][+FT_{R_{R_R}}]$
$T_{L_R} \rightarrow [-FT_{L_{R_L}}][+FT_{L_{R_R}}]$	$T_{R_{R_L}} \rightarrow [-F][+F]$
$T_{L_{R_L}} \rightarrow [-F][+F]$	$T_{R_{R_R}} \rightarrow [-F][+F]$
$T_{L_{R_R}} \rightarrow [-F][+F]$	

Definition 2. *Isomorphism on level X in rewriting rules. Rewriting rule R₁ and rewriting rule R₂ are isomorphic on depth X if they are homomorphic and their non-terminals are relatively isomorphic on depth X - 1. Isomorphic on level 0 indicates homomorphism.*

After defining the similarity between rules by homomorphism and isomorphism, we can classify all rules into different subsets where the rules in each subset has the same similarity relation. We will use the rule name as the class name. For example, we assign the terminal rewriting rule a class, “C₃ → null”. Assign a rule linked to two terminals, “C₂ → C₃C₃”, here C₃ is the terminal class. After classification, we obtain a context-free grammar set, which can be converted into an automata. After transforming the binary tree in Fig. 2 into the set of rewriting rules in Table 1, we can do the classification operation and get the results listed in Table 2.

1.3 Complexity for Classified Rules

The generating function [2] of a context free grammar is defined in the following paragraph.

Definition 3. *Generating function of a context free grammar.*

1. Assume that there are n classes of rules, {C₁, C₂, ..., C_n}, and the class C_i contains n_i rules. Let V_i ∈ {C₁, C₂, ..., C_n}, U_{ij} ∈ {R_{ij}, i = 1, 2, ..., n; j = 1, 2, ..., n_i}, and a_{ijk} ∈ {x : x = 1, 2, ..., n}, where k ∈ {1, 2} in binary

Table 2. Classification based on the similarity of rewriting rules. Several parameters for nodes and subtrees are attached to their symbols that will be used in the following formulas.

Classification of Rules Isomorphic Depth #2		
$(n = 10)$	Class #1 ($n_1 = 3$)	$n_{11} \quad a_{111}a_{112}$ $(1)C_1 \rightarrow C_1C_1$
		$n_{12} \quad a_{121}a_{122}$ $(1)C_1 \rightarrow C_4C_3$
		$n_{13} \quad a_{131}a_{132}$ $(1)C_1 \rightarrow C_4C_2$
Classification of Rules Isomorphic Depth #1		
	Class #2 ($n_2 = 1$)	$n_{21} \quad a_{211}a_{212}$ $(1)C_2 \rightarrow C_5C_7$
	Class #3 ($n_3 = 1$)	$n_{31} \quad a_{311}a_{312}$ $(1)C_3 \rightarrow C_7C_5$
	Class #4 ($n_4 = 1$)	$n_{41} \quad a_{411}a_{412}$ $(2)C_4 \rightarrow C_8C_6$
Classification of Rules Isomorphic Depth #0		
	Class #5 ($n_5 = 1$)	$n_{51} \quad a_{511}a_{512}$ $(2)C_5 \rightarrow C_9C_9$
	Class #6 ($n_6 = 1$)	$n_{61} \quad a_{611}a_{612}$ $(2)C_6 \rightarrow C_9C_{10}$
	Class #7 ($n_7 = 1$)	$n_{71} \quad a_{711}a_{712}$ $(2)C_7 \rightarrow C_{10}C_9$
	Class #8 ($n_8 = 1$)	$n_{81} \quad a_{811}a_{812}$ $(2)C_8 \rightarrow C_{10}C_{10}$
	Class #9 ($n_9 = 1$)	$n_{91} \quad a_{911}a_{912}$ $(8)C_9 \rightarrow \text{null}$
	Class #10 ($n_{10} = 1$)	$n_{10,1} \quad a_{10,11}a_{10,12}$ $(8)C_{10} \rightarrow \text{null}$

case. Each U_{ij} has the following form for a binary tree:

$$\begin{aligned} U_{i1} &\rightarrow V_{a_{i11}}V_{a_{i12}} \\ U_{i2} &\rightarrow V_{a_{i21}}V_{a_{i22}} \\ &\dots \rightarrow \dots \\ U_{in_i} &\rightarrow V_{a_{in_i1}}V_{a_{in_i2}}. \end{aligned}$$

2. The generating function of $V_i, V_i(z)$ has a form,

$$V_i(z) = \frac{\sum_{p=1}^{n_i} n_{ip} z V_{a_{ip1}}(z) V_{a_{ip2}}(z)}{\sum_{q=1}^{n_i} n_{iq}}.$$

If V_i does not have any non-terminal, we set $V_i(z) = 1$.

3. After formulating the generating function $V_i(z)$, we plan to find the largest value of z , z^{max} , where $V_1(z^{max})$ is convergent. Note that we will use V_1 to denote the function of the root node of the binary tree. After obtaining the largest value, z^{max} , of $V_1(z)$, we set $R = z^{max}$, where R is the radius of convergence of $V_1(z)$. Note R is a real value between zero and one. The complexity, K_0 , of the whole binary tree [1] is

$$K_0 = -\ln R.$$

4. Since computing the maximum value, z^{max} , directly is not feasible, we use iterations and region tests to accomplish the value. Rewrite the generating function in an iterative form,

$$V_i^m(z') = \frac{\sum_{p=1}^{n_i} n_{ip} z'^k V_{a_{ip1}}^{m-1}(z') V_{a_{ip2}}^{m-1}(z')}{\sum_{q=1}^{n_i} n_{iq}}; m = 1, 2, 3, \dots; \text{ and}$$

$$V_i^0(z') = 1.$$

5. The value of the function V_i at a specific z' can be calculated by iterations of the form. In each iteration, calculate the values from $V_i^0(z')$ to $V_i^m(z')$. When $V_i^{m-1}(z') = V_i^m(z')$ is satisfied for all classes, we stop the iteration. From experiences, we set $m = 200$ to simplify the calculations.
6. Now we can test whether $V_i(z')$ is convergent or divergent at a specific value z' . We use binary searching to test the values between 0 and 1. In each test, when $V_i(z')$ is convergent, we set a bigger value z' in the next test. When $V_i(z')$ is divergent, we set a smaller value z' in the next test. We expect that this test will approach the radius, $R = z^{max}$, closely.

2 Complexity of Encoded Text

We show how to compute the complexity of the text. A text sequence is first transformed into a binary string by a giving encoding method. One can directly set each character be an integer and obtain a binary string for the text. For example, use the integer indexes, 1 to 27, to represent the 26 alphabets plus the space character. A binary number with five bits is assigned to each alphabet. We use the term BIN to call this encoding method. This method is simple and doesn't apply any sophisticated encoding algorithm. A different method, Lempel-Ziv-Welch (LZW) [4], is also applied to encode the text. LZW is designed for lossless data compression and is a dictionary-based encoding. In LZW, when certain substring appears frequently in the text, it will be saved in the dictionary. These two methods will be used in this paper to accomplish the binary strings.

Before LZW processing, its dictionary contains all possible single characters of the text. LZW searches through the text sequence, successively, for a longer substring until it finds new one that is not in the dictionary. Whenever LZW finds a substring that is in the dictionary, its index is retrieved from the dictionary and this index will replace the substring's place in the encoded sequence. LZW will

add a new substring to the dictionary and attach a new index to the substring. The last character of the newly replaced substring will be used as the next starting character to scan for new substrings. Longer strings are saved, successively, in the dictionary and made available for subsequent encoding. When LZW operates on sequence with many repeated patterns, its compression efficiency is high.

For example, suppose there are only three characters “a”, “b”, “c” in the dictionary before we start searching. The indices, “1”, “2”, and “3”, are used to represent them respectively. Giving a text “abcabcabc”, the substrings, “ab”, “bc”, “ca”, “abc”, “cab”, will be saved in the dictionary successively with their new assigned indices, “4”, “5”, “6”, “7”, “8”. This string “abcabcabc” will be transformed into an array [1,2,3,4,6,5]. By using binary numbers, the array [1,2,3,4,6,5] can be transformed into a binary string ”001 010 011 100 110 101”.

Example

The article “The Declaration of Independence” is used as the text sequence. After removing all punctuations, the total number of characters is 7930. Apply the two encoding methods and obtain its two binary strings. Along the strings, every two connected binary bits “00”, “01”, “10”, and “11” are represented by four different small trees [5]. We calculate the complexity every 256 bits along each string. Fig. 3 shows that the complexity values of BIN are roughly fixed across the text sequence. The complexity values of LZW near the front end of the text are lower than those near the rear end of the text. These lower values reveal the encoding features of LZW. Since the LZW dictionary saves a lot of regular patterns in the front end and absorbs its regularities, there will be no such regular patterns in the rear end. We expect that the rewriting rules in L-system can capture the regularity in the sequence. A string section with high regularity has low complexity. So, the LZW string near the rear end becomes much random with high complexity. In BIN, the 256-bit text section that has the lowest complexity value (high regularity) is “...t all men are created equal, that they are endowed by ... ” .

3 Comparison with Topological Entropy

The topological entropy (TE) [6] is discussed and compared with the proposed method.

Topological Entropy

An information function $A_l(s)$ is defined as, $A_l(s) = |\{u : |u| = l \text{ and } u \text{ is a distinct substring in } s\}|$, where $A_l(s)$ represents the total number of distinct substrings with length l in the sequence s . The entropy is defined as,

$$H_l(s) = \frac{\log_k A_l(s)}{l},$$

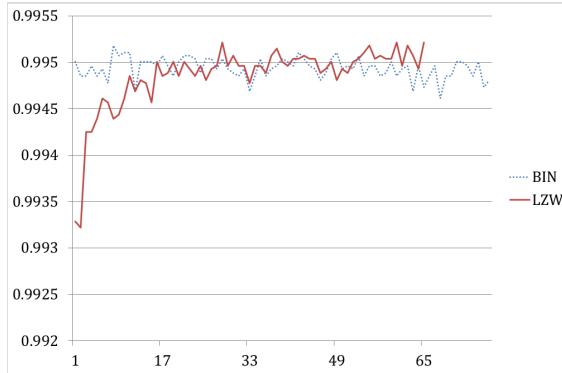


Fig. 3. The complexity of Declaration of Independence

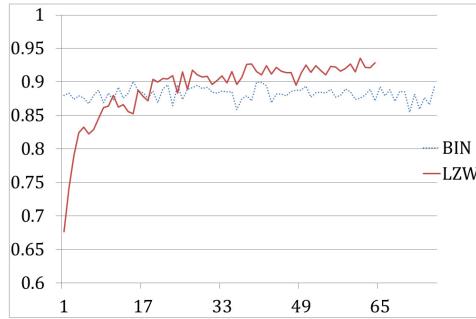


Fig. 4. The topological entropy of Declaration of Independence

where k is the size of distinct alphabets of the sequence. Since there are many values for l and this definition can't produce a single value of complexity for the whole sequence. A new definition for the topological entropy is in the following paragraph.

Definition 4. Let s be a finite sequence of length $|s|$ and k be the size of distinct alphabets, let l be the unique integer such that

$$k^l + l - 1 \leq |s| \leq k^{l+1} + (l + 1) - 1$$

We use $s_1^{k^l + l - 1}$ to represent the first $k^l + l - 1$ letters of s .

$$H_{TE}(s) := \frac{\log_k(A_l(s_1^{k^l + l - 1}))}{l}$$

where $A_l(s_1^{k^l + l - 1})$ is the number of distinct substrings with length l in sequence $s_1^{k^l + l - 1}$.

Fig. 4 shows the topological entropy values of the article “The Declaration of Independence”. Topological entropy focuses on only one subword length and computes its complexity. In contrast, linguistic complexity computes the complexity of all possible subword lengths. It uses much more computations than that of TE. The proposed method uses the binary tree to represent a binary sequence that has a length of power of 2. It reveals the structural information along a sequence.

Finally, we summarize several features of the proposed measure of complexity. The goal of this measure is different from that of the entropy developed for communication capacity. The rewriting rules in the proposed complexity can capture certain regularity in a sequence. To our knowledge, this measure is the only one that can sense the quasi-regular structure. It can be used in anomaly detection of a text by monitoring the outliers and statistics of its complexity values.

References

1. Kuich, W.: On the entropy of context-free languages. *Inf. Control* 16, 173–200 (1970)
2. Badii, R., Politi, A.: Complexity: Hierarchical structures and scaling in physics, p. 188. Cambridge University Press, Cambridge (1997)
3. Liou, C.-Y., Wu, T.-H., Lee, C.-Y.: Modeling complexity in musical rhythm. *Complexity* 15, 19–30 (2010)
4. Welch, T.A.: A technique for high-performance data compression. *IEEE Computer* 17, 8–19 (1984)
5. Liou, C.-Y., Tseng, S.-H., Cheng, W.-C., Tsai, H.-Y.: Structural complexity of DNA sequence. *Computational and Mathematical Methods in Medicine*, Article ID 628036 (2013)
6. Koslicki, D.: Topological entropy of DNA sequences. *Bioinformatics* 27, 1061–1067 (2011)

Compression in Molecular Simulation Datasets

Anand Kumar¹, Xingquan Zhu², Yi-Cheng Tu¹, and Sagar Pandit³

¹ Department of Computer Science and Engineering, University of South Florida,
Tampa, FL - 33620, U.S.A.
{akumar8,ytu}@cse.usf.edu

² Department of Computer Science and Engineering,
Florida Atlantic University, Boca Raton, FL - 33431, U.S.A.
xzhu3@fau.edu

³ Department of Physics,
University of South Florida, Tampa, FL - 33620, U.S.A.
pandit@cas.usf.edu

Abstract. In this paper, we present a compression framework, for molecular dynamics (MD) simulation data, which yields significant performance by combining the strength of principal component analysis (PCA) and discrete cosine transform (DCT). Though it is a lossy compression technique, the effect on analytics performed on decompressed data is very minimal. Compression ratio up to 13 is achieved with acceptable errors in results of analytical functions.

Keywords: molecular simulations, encoding, data compression, compression ratio, principal component analysis, discrete cosine transform.

1 Introduction

Many scientific disciplines, such as biochemistry, astronomy, and material sciences, are undergoing a radical change in research methodology from conducting “wet-bench” experiments to performing computer simulations. As a result, the particle simulations (PS) have seen tremendous efficiency improvements in the last decade. In PS, the system of interest (e.g., a protein and its environment) is studied as a collection of large number of basic components (e.g., atoms) whose behavior can be completely described by classical physics. Often such simulations generate spatio-temporal data that are in tera to peta bytes in size [1]. With increasing size of data, the problem of efficient storage and transfer persists. This paper addresses these issues in spatio-temporal data generated from PS applications by proposing effective data compression techniques.

1.1 Motivation and Contributions

The PS data is obtained from simulation results of a biological, physical or chemical phenomenon. In these systems, all individual atoms (we use atom and particle interchangeably) together represent large biological structures. Thus,

providing nano-scopic description of biological process. The simulation consists of measuring properties such as, 3D location of the atoms, velocity, charge, mass etc., at very small intervals of time (pico-seconds). Measurements of all atoms taken at a time instant, called snapshot (or *frame*), are stored on to computer disk. Considering the simulation for few microseconds, the data generated can easily reach terabytes. Since the simulation is generally done in large computer clusters, data needs to be transferred to a storage server. On the server end, data analytics face bottleneck due to limited I/O bandwidth. The above facts necessitate the compression of data for better utilization of the storage devices and network transfer bandwidth. The MD and the data management communities have been primarily focused on the high-performance computing, visualization and simple data management, thus left the problem of MD data compression inadequately addressed.

Traditional dictionary based approaches have inherent disadvantages in compressing particle simulation data: (1) compressed data size can still be large; and (2) whole data is scanned before starting compression. In addition, decompression becomes impossible if one or few data bytes are corrupted or damaged.

Contributions: The existing compression methods do not consider the temporal locality of the atoms for compression, which leaves a significant amount of redundancy in the compressed data. Our technique is designed to address these problems and meet the following four goals.

- High compression ratio (> 5), without noticeable errors is desirable, and dynamic error control to meet predefined requirements of compression quality.
- Error tolerance in compressed data and largely de-compressible. Even if some parts are corrupted, the errors are not propagated across many data frames.
- Access to random frames is allowed, without decompressing the whole data.
We achieve random access to a small group of frames.
- Balanced compression across different dimensions of the data.

In our framework, the MD data are first transformed, using PCA (Section 2.1), from the generic 3D coordinate space to another 3D eigen space, with the dimensions sorted in decreasing importance levels in capturing the variance of the atoms' movements. In the eigen space, the DCT is applied (Section 2.2) to achieve lossy compression across a window of consecutive frames. The lossy compression does not affect the results (Section 3) of the analytics that are often executed on molecular simulation data [2,3,4]. The combination of the PCA and DCT ensures that our framework can achieve aforementioned goals.

1.2 Related Work

Popular compression tools like WinZip and Gzip use dictionary based methods (such as Ziv-Lempel [5]). Statistical methods like Huffman encoding and arithmetic encoding [5] are used in compression of the multimedia data. These methods are either suitable for large text data or data that exhibit certain statistical properties. The resulting size of the compressed data can still be large.

Thus adding high cost to I/O and network transfers. There have been efforts to process approximate analytics using histograms [3], wavelets [4], and random sampling [6]. In these methods, the focus was mainly on the efficient data analysis while minimizing the I/O during execution time. A detailed survey of the traditional data compression techniques is provided by Salomon *et al.* [5].

A combination of different encoding-based compression techniques for molecular dynamics (MD) data is presented by Omletchenko *et al.* [2]. The technique utilizes spatial locality of the atoms using oct-tree index and space-filling curve (SFC), before encoding the data. Another approach to compression of trajectories and data management is provided by Essential Dynamics (ED) tool [7]. It is similar to compression of data using principal component analysis (PCA) [8], in which the trajectory pattern and the number of eigenvectors determine the error produced in the uncompressed data. These tools do not consider temporal locality of the atoms for compression. Hence, the achievable compression is limited. In molecular dynamics simulations, most atoms move along a trajectory of their own that changes very little over time. There is a huge scope for achieving high compression if temporal locality is considered. We attempt to achieve better compression in our approach by considering these features.

2 Compression Framework

In this section, we discuss the the basic framework of compression along with the encoding techniques used. Fig. 1 shows the framework of the proposed compression method. Our main theme is to employ PCA to transform data to another space, on which lossy compression can be achieved using DCT. Due to space limitations we omit some details of the modules.

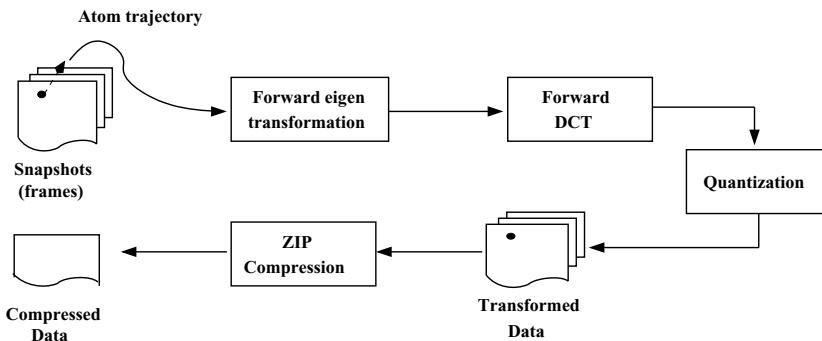


Fig. 1. The framework for compression of particle simulation data

2.1 Transforming Data to Eigen Space

The measurement values captured during molecular simulation change over time as the data is stored after each snapshot. The measurements depict changes in

certain properties of the atoms in the simulation. As the MD data is of large volume, we compress a group of snapshots (frames) called window. The number of frames in a window can be controlled depending on the availability of computational resources and the level of errors we are willing to tolerate in the final data. We apply orthogonal linear transformation to the data using principal component analysis (PCA) [8]. For each atom (as in Fig. 2) we collect its locations across a number of adjacent frames, f_1, f_2, \dots, f_n , each of which contributes a 3D location (l_x, l_y, l_z) . PCA transforms data to a new coordinate system (l_p, l_q, l_r) such that the greatest variance is observed on the first coordinate l_p after projection of the data. The coordinates are ordered in decreasing order of the variance of the projection, which is obtained by ordering the eigen values and corresponding eigenvectors of the data. The transformation \mathbf{F} of given data \mathbf{D} , with eigenvectors \mathbf{E} in decreasing order of eigen values, using PCA is given by equation (1).

$$\mathbf{F}^T = \mathbf{D}^T \mathbf{E} \quad (1)$$

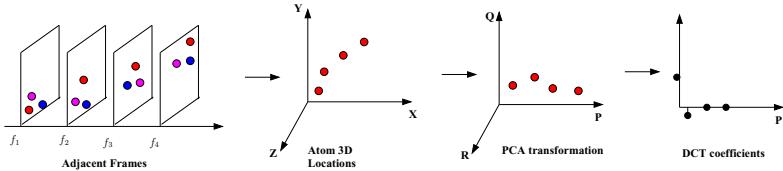


Fig. 2. Intuition behind compression: Performing principal component analysis (PCA) followed by discrete cosine transform (DCT) on data

The eigen values and eigenvectors can be found from the covariance matrix of the data in the window. The original data \mathbf{D} can be obtained back from the transformation using equation (2).

$$\mathbf{D}^T = \mathbf{E}^{-1} \mathbf{F} \quad (2)$$

The eigen space transformation aligns the data along eigenvectors based on the variance observed in the data. By transforming the data to the new coordinate system (P, Q, R) , we can pick few dimensions that represent the complete data (compressed) while maintaining low errors (in decompressed data). This is lossy compression. In our approach we transform the data into new coordinates to obtain the different directions of variance. These directions are then compressed independently using DCT. Each dimension uses different DCT parameters based on the degree of variance for compression. This gives data with minimal loss to achieve better performance of the DCT compression (Section 2.2).

2.2 Computing DCT of Transformed Data

The eigen analysis step transforms data into new dimensions [8]. We encode every dimension (attribute/measure) of the atoms separately. However, the eigen

space transformation is applied to 3D trajectories of each atom separately. Finally, we apply the DCT on each dimension P, Q, R (after dropping unnecessary dimensions) of the transformed data. By transforming the data to the new coordinate system, we are trying to reduce the errors in all dimensions that may occur in performing the DCT step. This makes recovering DCT coefficients more accurate during decompression, as the coefficients in the beginning contain more accurate information.

The atom localization property is utilized by applying the DCT to a constant number of adjacent frames, say N . Given a window of size N (frames), we compress single measurement of every atom individually. The trajectory of length N is transformed using DCT. Such transformation is applied to trajectories of all atoms in the window. Now, the DCT data \mathbf{C} is quantized to reduce the number of bits required to store the coefficients. The coefficients can be eliminated by assigning very low weights to the high frequency components. After all the data is processed, we apply ZIP compression on the DCT data. As a lossless compression method, the ZIP compression removes the redundant information present in the data. This gives the final compressed data. The combination of the PCA and DCT in the framework ensures that (1) compression is balanced across all dimensions, (2) error can be controlled dynamically; and (3) any portion of data can be accessed at random.

3 Experimental Results and Discussion

The proposed compression technique is tested on real molecular dynamics datasets. We used two different datasets, as shown in Table 1, for the experiments. These data sets were obtained from snapshots of real molecular simulations. The complete data set had around 600 frames or snapshots. The measurements stored in the data are: x, y and z coordinates of the atoms, charge and mass measured during the simulation.

Table 1. Compression of MD simulation datasets, using PCA and DCT ($W = 128$)

Data set	Objects	Data Size	Compressed	Ratio	RMSE (Å)
Protein	286,849	3.8 GB	293.45 MB	13.26	0.14
Collagen Fiber	891,272	6.4 GB	535.86 MB	12.23	0.09

The effect of compression on quality of the data is measured as the error in final decompressed data. The error is measured in terms of root mean square error (RMSE; using equation in Table 2). Distance between original data frame F_u and the compressed data frame F_c (we call the data obtained after *decompression* as compressed) is used to compute the error. N_a is the number of atoms in every frame, and N_f is the total number of frames in the data set. We computed the RMSE on locations of objects present in the system.

Table 2. Equations required in experiments and analysis of results

$Error = \sqrt{\frac{1}{N_a \times N_f} \sum_{i=1}^{N_f} \sum_{j=1}^{N_a} \{F_c - F_u\}}$	$MSD(t) = \frac{1}{N} \sum_{i=1}^N \mathbf{r}_i(t) - \mathbf{r}_i(0) ^2$
$MC(t) = \frac{\sum_{i=1}^N m_i \mathbf{r}_i(t)}{\sum_{i=1}^N m_i}$	$RDF(r) = \frac{N(r)}{4\pi r^2 \sigma r \rho}$

The results shown in Table 1 explain the performance of the proposed compression technique. In this method we are able to achieve high compression ratio. The results shown in Table 1 are obtained with a window size of 128. Effect of window size on the compression ratio of the data is shown in Table 3.

Table 3. Effect of window size on Protein data compression

Window Size	Compressed Size	Compression Ratio	RMSE (Å)
008	987.61 MB	3.94	0.25
016	603.29 MB	6.45	0.23
032	432.36 MB	9.00	0.21
064	353.75 MB	11.00	0.18
128	293.45 MB	13.26	0.14
256	386.03 MB	10.08	0.09
512	407.46 MB	9.55	0.05

The compression ratio directly depends on the amount of error that can be allowed to appear in the final data. The errors can be controlled by adjusting coefficients that are retained (as explained in section 2) after the DCT step. The error should increase with the compression ratio, for a fixed window size. This effect is observed in our dataset also. The plot shown in Fig. 3(a) presents this relation between the compression ratio and error. The RMSE increases as the compression ratio is increased. Our PCA and DCT based compression technique gives better compression ratio while maintaining low errors in the compressed data. This method is best suited for compression of very large volumes of data, which are common in the scientific databases.

Data Analysis Results

The effect of compression on the data set can also be measured using the results of analytical queries. We applied some queries, interesting to researchers, on the decompressed data to measure this effect.

Mean Square Displacement (MSD): Consider a system with N particles. Let $\mathbf{r}_i(t)$ be the location of particle i , of mass m_i , at time instant t . The MSD information biophysicists ask frequently is formulated by equation given in Table 2. The proposed compression technique introduces very small error in the MSD. The

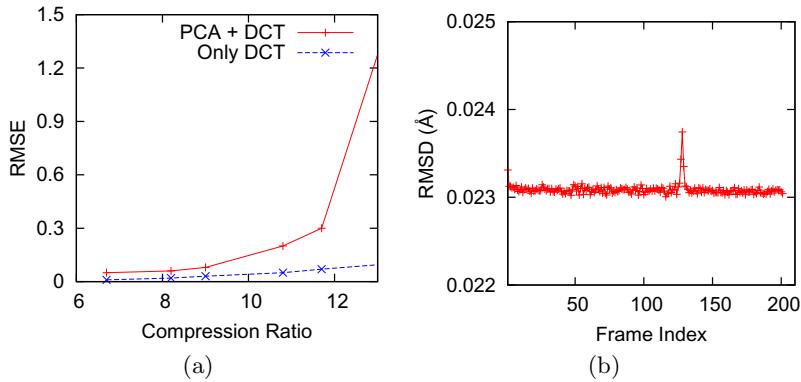


Fig. 3. (a) Average RMSE (in Å) plotted against compression ratio. Window size of 128 is chosen 3(b) RMSD (in Å) between original and decompressed frames.

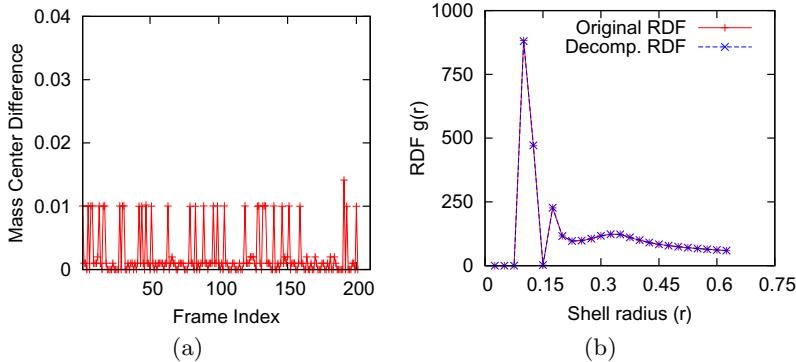


Fig. 4. Effect of compression: 4(a) On mass centers (in Å). 4(b) Radial distribution function (RDF). Shell resolution $\delta r = 0.025\text{Å}$ and window size of 128 frames was used.

plot of Fig. 3(b) shows the root mean square displacement of the compressed frame from the original frame. A very small peak in the plot indicates the start of a new window used for the compression. Window size of 128 frames was used for this particular experiment.

Mass Center (MC): Another information of interest to biophysicists is the mass center (MC, Table 2) of the system at any given time instant. Fig. 4(a) shows the effect on mass center, $MC(t)$, of the particle space at a given time instant t . It can be seen that displacement in the mass center is very small. Small fluctuations in the plot are the effects of DCT coefficient quantization. This shows that the proposed method performs well in compression and decompression of the simulation data.

Radial Distribution Function (RDF): Main motivation behind the proposed compression approach is to demonstrate the minimal effect on analytical queries.

We observed this behavior in the experimental results on RDF [9]. The RDF is defined by equation sown in Table 2, where $N(r)$ is the number of atoms in the shell between r and $r + \delta r$ around any particle, ρ is the average density of particles in the whole system, and $4\pi r^2 \delta r$ is the volume of the shell. The RDF can be viewed as a normalized spatial distance histogram (SDH). The SDH is a fundamental tool in the validation and analysis of particle simulation data. Fig. 4(b) shows that RDF is not much affected by compression. For shell resolution $r > 0.025\text{\AA}$, the RDF values in the compressed and decompressed data are almost same. This shows that the error introduced by compression on RDF of the frames is minimal. Hence, the proposed technique is suited for compression of the molecular simulation data.

4 Conclusions and Future Work

A compression technique for MD data, using PCA and DCT, that can achieve compression ratio of about 13 is presented in this paper. Temporal locality is exploited to achieve good compression that can be used to satisfy the I/O and network bandwidth requirements. Some of the problems that need to be addressed in future are: (1) analysis on compressed data directly; (2) efficient encoding technique, to utilize the correlation between trajectories; and (3) data mining to find patterns of interest in the trajectories. The findings that we reported in this paper will definitely lead to abundant research efforts.

Acknowledgments. This project (R01GM086707) is supported by the National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH), USA.

References

1. Etten, W.V.: Managing data from next-gen sequencing. *Genetic Engineering and Biotechnology News* 28(8) (2008)
2. Omelchenko, A., et al.: Scalable i/o of large-scale molecular dynamics simulations: A data-compression algorithm. *Computer Physics Comm.* 131(1-2), 78–85 (2000)
3. Ioannidis, Y.E., Poosala, V.: Histogram-based approximation of set-valued query-answers. In: *Procs. of VLDB*, pp. 174–185 (1999)
4. Chakrabarti, K., Garofalakis, M., Rastogi, R., Shim, K.: Approximate query processing using wavelets. *The VLDB Journal* 10(2-3), 199–223 (2001)
5. Salomon, D.: *Data Compression: The Complete Reference*. Springer (2004)
6. Cochran, W.G.: *Sampling Techniques*, 3rd edn. John Wiley and Sons (1977)
7. Meyer, T., et al.: Essential dynamics: A tool for efficient trajectory compression and management. *Journal of Chemical Theory Computation* 2(2), 251–258 (2006)
8. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. Wiley-Interscience Publication (2000)
9. Bamdada, M., et al.: A new expression for radial distribution function and infinite shear modulus of lennard-jones fluids. *Chemical Physics* 325, 554–562 (2006)

Online-Learning Structural Appearance Model for Robust Visual Tracking

Min Yang, Mingtao Pei, Yuwei Wu, Bo Ma, and Yunde Jia

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology,
Beijing 100081, P.R. China

{yangminbit, peimt, wuyuwei, bma000, jiayunde}@bit.edu.cn

Abstract. The main challenge of robust visual tracking comes from the difficulty in designing an adaptive appearance model to account for appearance variations. Existing tracking algorithms often build an representation for the tracked object, and perform self-updating of the object representation with examples from recently tracking results. Slight inaccuracies in the tracker can degrade the appearance models. In this paper, we propose a robust tracking method with an online-learning structural appearance model based on local sparse coding and online metric learning. Our appearance model employs structural feature pooling over the local sparse codes of an object region to obtain a robust object representation. Tracking is then formulated as seeking for the most similar candidate within a Bayesian inference framework where the distance metric used for similarity measurement is learned in an online manner to match the varying object appearances. Both qualitative and quantitative evaluations on various challenging image sequences demonstrate that the proposed algorithm outperforms the state-of-the-art methods.

Keywords: Visual tracking, appearance modeling, sparse coding, online metric learning.

1 Introduction

Appearance modeling is a critical prerequisite for successful visual tracking. An appearance model generally consists of two modules: object representation which captures the visual characteristics of an object and appearance matching scheme that measures the similarity between observed samples and the model. Due to appearance variations caused by background clutters, object deformation, illumination changes and occlusions *etc*, designing a robust appearance model is a challenging task.

Recently, sparse representation based appearance modeling has received considerable attention in the visual tracking community [12,10,15,4,19]. The pioneer work introduced by Mei and Ling [12] models the object appearance as a sparse linear combination of object and trivial templates via ℓ_1 minimization. Wang *et al.* [15] employed subspace learning method to construct and update the object

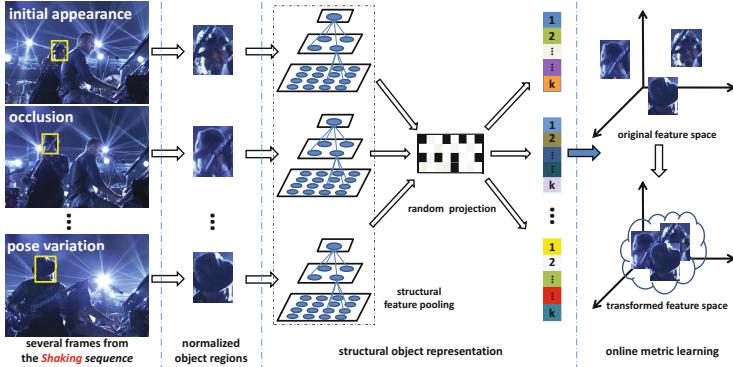


Fig. 1. Motivation of our work. The tracked object is represented by using a structural representation strategy, and the varying object appearances caused by occlusions and pose variations *etc.* can be successfully matched via online metric learning.

templates used for sparse representation. However, only the holistic information of the object is used in these methods, which makes it difficult to handle drastic view or pose variations. Zhong *et al.* [19] presented a sparsity-based collaborative appearance model which exploits the advantages of both holistic and local information. Jia *et al.* [4] introduced an alignment-pooling method across the sparse codes of local patches to improve the accuracy of location estimation. Motivated by the success of sparse representation for object tracking, one aspect we focus on is to design an effective object representation containing local and spatial information using sparse coding.

Most of appearance models based on sparse representation measure the similarity between the candidates and the model by reconstruction errors [12,15,19]. However, when the object undergoes significant appearance variations, the true candidate of the object might have large reconstruction errors, leading to ambiguity of the tracker. In fact, the magnitude of the reconstruction errors depend largely on the dictionary which should be updated in an online manner to account for the varying appearances. Nevertheless, straightforward updating of the dictionary with newly obtained results is prone to potential drift because of the accumulated errors. To keep the flexibility, our another emphasis is placed on seeking for a suitable feature space via online metric learning to match the varying appearances rather than updating the dictionary to ensure the correspondence between minimal reconstruction error and true object location.

Metric learning has been introduced to object tracking by several methods [5,17,9]. Jiang *et al.* [5] integrated neighborhood component analysis (NCA) into kernel-based tracking framework to improve the tracking performance. Wang *et al.* [17] formulated appearance modeling and motion estimation into a unified framework based on metric learning. However, methods proposed in [5,17] only use simple representation strategies, *e.g.*, color histogram, to represent the object appearance, thus are sensitive to significant appearance variations. Furthermore, these methods learn the distance metric in an off-line manner, which often leads

to expensive computation. Li *et al.* [9] presented a non-sparse linear representation with the learned Mahalanobis distance metric for visual tracking. This method only utilizes the holistic information and ignores the trivial template, resulting in failure of tackling occlusions.

Considering the two objectives mentioned above, we develop an online-learning structural appearance model for robust visual tracking. More specifically, we sample several image patches inside an object region using overlapped sliding windows, and employ a structural feature pooling process to concatenate the sparse codes of these patches into a structural representation of the object region. This structural representation captures local and spatial information of the image patches, followed by a very sparse random projection to generate a low-dimensional compact representation. An online metric learning algorithm is then advocated to get a discriminative and adaptive metric for appearance matching. The learned metric makes the different appearances of the object close to each other, and separates the object from the background simultaneously. The main components of our tracking method are depicted in Fig. 1. Numerous experiments and evaluations on challenging video sequences demonstrate that our method outperforms several state-of-the-art trackers.

2 Structural Object Representation

Given a normalized object region \mathbf{P} , we first sample N local patches inside the region using overlapped sliding windows. Each patch representing one fixed part of an object is then converted to a d -dimensional vector $\mathbf{p}_j \in \mathbb{R}^{d \times 1}$. Therefore, the complete structure of the object can be represented by concatenating all these patches together, *i.e.*, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N] \in \mathbb{R}^{d \times N}$. Let \mathbf{D} be a dictionary (or codebook) with n entries, $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n] \in \mathbb{R}^{d \times n}$, then each patch \mathbf{p}_j can be converted into a n -dimensional code using sparse coding.

2.1 Locality-Constrained Linear Coding

In visual tracking applications, similarity is more essential than sparsity [10]. The regularization term of ℓ_1 norm in traditional sparse coding scheme is not smooth, which leads to the loss of correlations between codes even though for similar patches. Hence, we employ locality-constrained linear coding (LLC) [16] to preserve the similarity between image patches. Specifically, the LLC code $\mathbf{x}_j \in \mathbb{R}^{n \times 1}$ corresponding to $\mathbf{p}_j \in \mathbb{R}^{d \times 1}$ is computed by

$$\begin{aligned} & \min_{\mathbf{x}_j} \|\mathbf{p}_j - \mathbf{D}\mathbf{x}_j\|_2^2 + \lambda \|\mathbf{e}_j \odot \mathbf{x}_j\|_2, \\ & \text{s.t. } \mathbf{1}^\top \mathbf{x}_j = 1 \end{aligned} \quad (1)$$

where \odot denotes element-wise multiplication, and \mathbf{e}_j is the Euclidean distance vector between \mathbf{p}_j and all basis vectors in \mathbf{D} . Note that the LLC code in Eqn. 1 is not ℓ_0 norm sparse, but is sparse in the sense that the solution only has few significant values. LLC actually selects a set of local basis vectors for \mathbf{p}_j to form a local coordinate system. In this work, we use a fast approximation of LLC [16] to efficiently obtain the sparse codes of image patches.

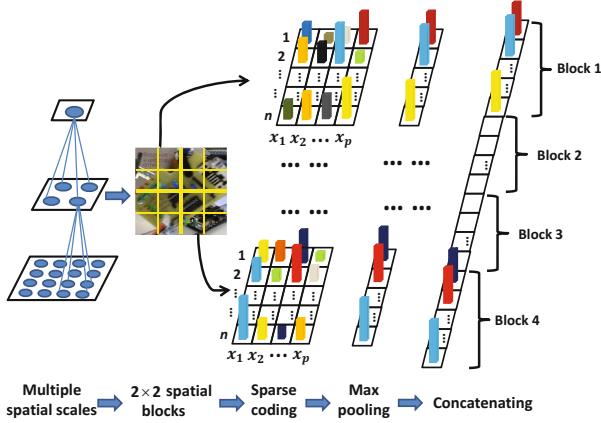


Fig. 2. Illustration of our structural feature pooling process. We show three spatial scales and take 2×2 as an example. In order to illustrate the pooling process clearly, we use 8×8 non-overlapped sliding window to obtain $4 \times 4 = 16$ images patches within the region (the object region is normalized into 32×32). Each of the four spatial blocks (*i.e.*, Block 1, \dots , Block 4) contains $p = 4$ images patches. The higher cylinders in the figure represent the larger values of LLC codes.

2.2 Feature Pooling

Denote the LLC codes of an object as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{n \times N}$, the feature pooling function on LLC codes is defined as $\boldsymbol{\alpha} = \xi(\mathbf{X})$, where $\boldsymbol{\alpha} \in \mathbb{R}^{n \times 1}$ and $\xi(\cdot)$ is defined on each row of \mathbf{X} . We use the max pooling process to acquire a middle level representation, which is well established with biophysical evidence in visual cortex and has been shown to be effective for image representation [18]. Therefore, the i -th element of $\boldsymbol{\alpha}$ is given by

$$\alpha_i = \max \{|\mathbf{x}_{i,1}|, |\mathbf{x}_{i,2}|, \dots, |\mathbf{x}_{i,N}|\}, \quad (2)$$

where $\mathbf{x}_{i,j}$ is the element at i -th row and j -th column of \mathbf{X} . In this case, $\boldsymbol{\alpha}$ is a global pooled feature, since the pooling function is measured on the whole image patches, discarding the spatial information of the local patches.

In order to capture local and spatial information, we construct a spatial pyramid for the object region and do max pooling on multiple spatial scale. Suppose the object region \mathbf{P} is partitioned into $\sigma \times \sigma$ non-overlapped spatial blocks $\{\Delta_b\}_{b=1}^{\sigma^2}$ on the spatial scale σ , the LLC codes \mathbf{X} are accordingly divided into $\sigma \times \sigma$ subsets $\{\mathbf{X}_{\Delta_b}\}_{b=1}^{\sigma^2}$. The corresponding pooled features are denoted as $\{\boldsymbol{\alpha}_b = \xi(\mathbf{X}_{\Delta_b})\}_{b=1}^{\sigma^2}$. Concatenating the pooled features created from each subset of the LLC codes on various spatial scales, we can acquire a structural representation \mathbf{Z}^* of the object region,

$$\mathbf{Z}^* = [\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \dots, \boldsymbol{\alpha}_\nu^\top]^\top, \quad (3)$$

where $\mathbf{Z}^* \in \mathbb{R}^m$, $m = n \times \nu$ and ν is the total number of spatial blocks on all spatial scales. Fig. 2 illustrates our structural feature pooling process, in which we show three spatial scales and take 2×2 as an example to demonstrate the feature pooling on a specific spatial scale.

2.3 Dimensionality Reduction

The feature vector \mathbf{Z}^* described in Sect. 2.2 is usually high dimensional. We use the random projection to embed the $\mathbf{Z}^* \in \mathbb{R}^m$ into a low-dimensional subspace $\mathbf{Z} = \mathbf{R}\mathbf{Z}^*$, where $\mathbf{R} \in \mathbb{R}^{m \times k}$ is a random projection matrix and $\mathbf{Z} \in \mathbb{R}^k$. A *very sparse random projection* [8] is introduced to help find effective subspace for the original data. The entries of the random projection matrix \mathbf{R} are defined as

$$r_{ij} = \sqrt{q} \times \begin{cases} +1 & \text{with probability } 1/2q \\ 0 & \text{with probability } 1 - 1/q \\ -1 & \text{with probability } 1/2q \end{cases}, \quad (4)$$

where $q = \sqrt{m}$. Compared with the traditional dimensionality reduction methods, e.g., PCA, the random projection matrix defined by Eq. 4 is independent with the original data, and hence is suitable for our framework.

3 Online Metric Learning

Given an object template $\mathbf{Z}_T \in \mathbb{R}^k$ created from the first frame and a candidate $\mathbf{Z}_C \in \mathbb{R}^k$ in the current frame, the Mahalanobis distance between \mathbf{Z}_T and \mathbf{Z}_C is defined as

$$\mathcal{D}_{\mathbf{M}}(\mathbf{Z}_T, \mathbf{Z}_C) = (\mathbf{Z}_T - \mathbf{Z}_C)^{\top} \mathbf{M} (\mathbf{Z}_T - \mathbf{Z}_C), \quad (5)$$

where $\mathbf{M} \in \mathbb{R}^{k \times k}$ is required to be a symmetric positive semi-definite matrix. In this work, the matrix \mathbf{M} is adaptively obtained by an online metric learning method, and the object template remains fixed during tracking.

We first describe the training examples collection mechanism in our algorithm. Once the object is located, we sample a set of image regions from a small neighborhood around the object location, and label the feature vectors of these regions as positive examples. Similarly, the negative examples are composed of the feature vectors of the image regions far away from the object location.

Following the method in [14], we encode a tuple used for online metric learning as (u, v, l) , where (u, v) is a example pair and l is the label which equals $+1$ if u and v are considered similar and -1 otherwise. Given a tuple set, a margin constraint is that the distances between all pairs of dissimilar examples are greater than the distances between all pairs of similar examples at least γ . Alternatively, there exists a threshold δ which is subject to the rule:

$$\begin{cases} \mathcal{D}_{\mathbf{M}}(u, v) \leq \delta - \gamma/2 & \forall (u, v, l) : l = +1, \\ \mathcal{D}_{\mathbf{M}}(u, v) \geq \delta + \gamma/2 & \forall (u, v, l) : l = -1. \end{cases} \quad (6)$$

We set γ to be 2, and rewrite the constraint as

$$l \cdot (\delta - \mathcal{D}_M(u, v)) \geq 1. \quad (7)$$

During online learning, at each time step τ , we get a tuple (u_τ, v_τ, l_τ) and calculate the distance $\mathcal{D}_{M_\tau}(u_\tau, v_\tau)$ between the two examples according to the current metric M_τ . After getting the prediction label $\hat{l}_\tau = sign(\mathcal{D}_{M_\tau}(u_\tau, v_\tau) < \delta_\tau)$, we can compute a loss if there is a discrepancy between \hat{l}_τ and l_τ . The loss function is given by

$$\phi_\tau(M, \delta) = \max \left\{ 0, l_\tau (\mathcal{D}_{M_\tau}(u_\tau, v_\tau) - \delta_\tau) + 1 \right\}. \quad (8)$$

The goal of the online algorithm is to minimize the cumulative loss in Eq. (8). According to [14], this can be done by updating the matrix M_τ and the threshold δ_τ using two successive projections

$$\begin{aligned} (M_{\hat{\tau}}, \delta_{\hat{\tau}}) &= \mathcal{P}_{C_\tau}(M_\tau, \delta_\tau), \\ (M_{\tau+1}, \delta_{\tau+1}) &= \mathcal{P}_{C_a}(M_{\hat{\tau}}, \delta_{\hat{\tau}}), \end{aligned} \quad (9)$$

where $\mathcal{P}_C(v)$ indicates the orthogonal projection from vector v to a closed convex set C , $C_\tau = \{(M, \delta) : \phi_\tau(M, \delta) = 0\}$ is the set of all (M_τ, δ_τ) pairs which attain zero loss on the example (u_τ, v_τ, l_τ) , and $C_a = \{(M, \delta) : M \succeq 0, \delta \geq 1\}$ is the set of all admissible (M_τ, δ_τ) pairs.

4 Proposed Tracking Algorithm

Object tracking can be considered as a Bayesian inference task in a Markov model with hidden state variables. Given the observed image set $\mathcal{O}_{1:t} = \{\mathbf{o}_1, \dots, \mathbf{o}_t\}$ up to time t , the optimal state \mathbf{s}_t of the tracked object can be estimated by Bayesian theorem

$$p(\mathbf{s}_t | \mathcal{O}_{1:t}) \propto p(\mathbf{o}_t | \mathbf{s}_t) \int p(\mathbf{s}_t | \mathbf{s}_{t-1}) p(\mathbf{s}_{t-1} | \mathcal{O}_{1:t-1}) d\mathbf{s}_{t-1}. \quad (10)$$

This inference is governed by the dynamic model $p(\mathbf{s}_t | \mathbf{s}_{t-1})$ and the observation model $p(\mathbf{o}_t | \mathbf{s}_t)$. A particle filter [3] is used to approximate the posterior $p(\mathbf{s}_t | \mathcal{O}_{1:t})$ by a finite set of N_s samples $\{\mathbf{s}_t^i\}_{i=1}^{N_s}$ with importance weights $\{\omega_t^i\}_{i=1}^{N_s}$. We apply an affine image warp to model the object motion between two consecutive frames. The dynamic model $p(\mathbf{s}_t | \mathbf{s}_{t-1})$ is modeled by Brownian motion, i.e., $p(\mathbf{s}_t | \mathbf{s}_{t-1}) = \mathcal{N}(\mathbf{s}_t; \mathbf{s}_{t-1}, \Sigma)$, where Σ is a diagonal covariance matrix.

Given a candidate sample \mathbf{s}_t^i , the region of interest can be extracted from the observed image \mathbf{o}_t by applying an affine transformation using \mathbf{s}_t^i as parameters. Then the observation likelihood of the candidate \mathbf{s}_t^i is computed by

$$p(\mathbf{o}_t | \mathbf{s}_t^i) \propto \exp(-\mathcal{D}_M(\mathbf{Z}_T, \mathbf{Z}_C^i)), \quad (11)$$

where \mathbf{Z}_T and \mathbf{Z}_C^i are feature vectors of the template and the candidate \mathbf{s}_t^i , respectively. For the tracking at time t , the candidate with the maximum observation likelihood is chosen as the tracking result.

5 Experimental Results

We run our tracking algorithm on twelve public challenging video sequences, and only gray scale information is used for our experiments. The challenges of these videos include heavy occlusion, illumination changes, pose variations, motion blur, scale variations and complex backgrounds. Our tracker is compared against seven state-of-the-art tracking algorithms denoted as Frag [1], IVT [13], ℓ_1 [12], TLD [6], MIL [2], VTD [7] and SCM [19], respectively. We use the source codes provided by the authors with the same initialization and their default parameters. Since the trackers except Frag involve randomness, we run them 5 times and report the average result for each sequence.

5.1 Implementation Details

We resize the object image to 32×32 pixels and extract overlapped 8×8 patches within the object region with 2 pixels as the step length. Performing k -means clustering algorithm on the patches extracted from first frame, the number of dictionary entries n is set to 100. The multi-scale max pooling is performed on three spatial scales 1×1 , 2×2 and 3×3 blocks, resulting in a 1400-dimensional feature vector. As discussed in [11], we set the random projection dimensionality $k = 400 \approx 1400/3$. Given the object location at the current frame, 20 positive examples and 15 negative examples are collected for online metric learning. As a trade-off between computational efficiency and effectiveness, the metric matrix M is updated every 10 frames. *The parameters are fixed for all sequences.*

5.2 Quantitative Evaluation

We use the center location error as well as the overlap rate for quantitative evaluations. Center location error is the per-frame distance (in pixels) between the center of the tracking result and that of the ground truth. Overlap rate is defined as $\frac{\text{area}(R_T \cap R_G)}{\text{area}(R_T \cup R_G)}$, where R_T is the bounding box of tracking result and R_G denotes the ground truth. Table 1 and Table 2 summarize the average center location errors and the average overlap rates, respectively. Note that the TLD tracker does not give tracking result when occlusions occur and the object is need to be re-detected. Thus, we only show the center location errors for the sequences that the TLD can keep track all the time. Overall, the proposed tracker performs favorably against the state-of-the-art algorithms.

To validate the effectiveness of online metric learning, we present the tracking results using a fixed distance metric learned by using the training examples extracted from the first frame (denoted as **Ours w/o ML**) in Table 1 and Table 2. The results show that the online metric learning mechanism provides an effective way to account for appearance variations of the object, and facilities appearance updating and object tracking. More interestingly, **Ours w/o ML** also performs well in some sequences, which demonstrate the effectiveness of the proposed structural object representation.

Table 1. Average center location error (in pixels). **Bold** fonts indicate the best performance while the *italic* fonts indicate the second best ones.

	Frag	IVT	ℓ_1	TLD	MIL	VTD	SCM	Ours w/o ML	Ours
Bird2	25.3	102.1	125.0	—	13.2	54.8	11.9	<i>11.0</i>	6.7
Bolt	166.6	212.0	166.9	—	7.0	95.0	94.4	39.6	<i>7.3</i>
Surfer	75.1	125.7	119.3	—	85.1	53.5	35.9	<i>15.0</i>	11.6
Car6	49.5	53.1	5.2	—	89.8	53.4	841.0	27.0	4.8
Caviar2	5.7	8.5	54.6	7.2	70.0	6.4	2.7	<i>2.8</i>	2.9
Woman	<i>119.2</i>	246.6	154.8	—	121.2	135.6	121.2	122.7	5.6
David	69.6	6.0	57.7	6.4	28.9	27.8	5.1	8.3	<i>5.2</i>
Shaking	118.5	124.5	56.6	—	19.7	6.3	8.6	8.8	7.7
Singer2	37.9	88.5	63.0	—	59.4	<i>18.1</i>	53.7	177.8	10.5
Panda	100.0	95.6	82.6	—	42.9	78.2	3.7	2.9	3.1
Jumping	8.3	4.4	44.0	4.1	39.5	73.0	4.1	113.0	3.6
Board	85.8	179.3	197.4	150.3	65.5	78.9	<i>20.4</i>	27.6	<i>15.0</i>

Table 2. Average overlap rate (%). **Bold** fonts indicate the best performance while the *italic* fonts indicate the second best ones.

	Frag	IVT	ℓ_1	TLD	MIL	VTD	SCM	Ours w/o ML	Ours
Bird2	47.5	10.6	8.4	17.3	67.8	15.5	69.5	71.4	79.6
Bolt	1.3	1.1	15.4	1.1	71.9	1.8	13.5	37.0	<i>71.3</i>
Surfer	12.8	5.3	5.0	35.4	16.6	20.3	30.1	56.1	64.2
Car6	54.0	40.3	78.7	76.8	14.8	51.5	3.2	63.8	80.2
Caviar2	53.6	45.7	33.5	68.1	23.2	61.1	80.2	81.1	81.7
Woman	16.3	<i>16.6</i>	5.4	10.6	15.5	14.9	15.2	16.5	71.8
David	29.9	64.8	29.6	55.6	47.7	49.3	42.5	<i>73.7</i>	83.9
Shaking	21.3	3.0	14.2	12.1	58.7	73.1	71.9	70.7	72.3
Singer2	50.8	23.3	23.9	9.7	24.0	67.8	33.0	8.7	72.9
Panda	32.1	9.3	1.4	58.7	45.1	37.2	63.5	62.0	64.4
Jumping	58.6	70.6	14.4	65.5	20.9	11.5	72.2	10.6	73.5
Board	58.5	12.6	10.0	16.8	46.6	38.8	76.2	67.2	<i>73.1</i>

5.3 Qualitative Evaluation

Several screenshots of the visual tracking results on the twelve sequences are illustrated in Fig. 3. We give a qualitative evaluation of the tracking results in four different ways as follows.

Pose Variation. In the *Bird2*, *Bolt* and *Surfer* sequences, the object appearances change drastically due to significant pose variations. We can see that only our method tracks the objects successfully in all these three sequences. Other evaluated algorithms except the MIL and SCM methods fail when the objects start to change their pose (*e.g.*, *Bird2* #50 and *Bolt* #20). In the *surfer* sequence, the MIL and SCM methods gradually drift away when there is severe occlusion and large scale change of the object. Our method adaptively cope with appearance variations via online metric learning, thus provide more accurate and consistent tracking results.

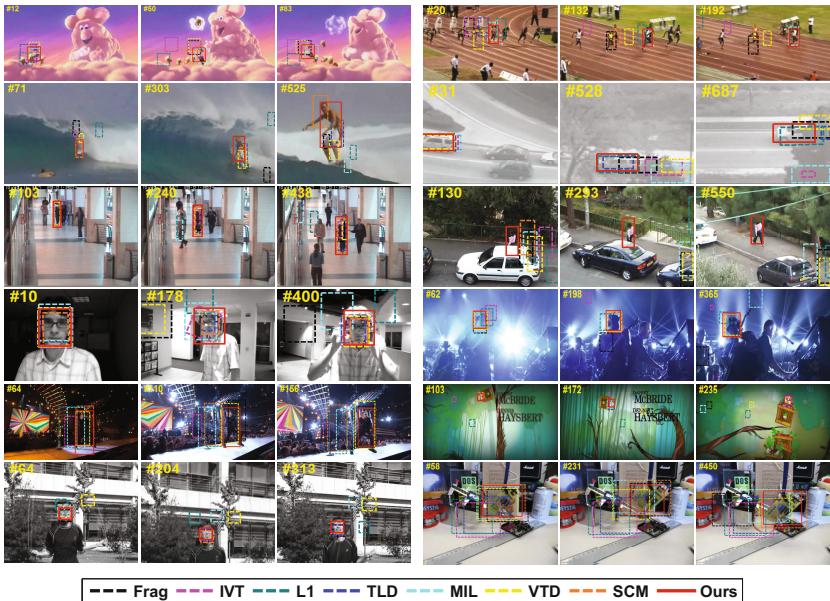


Fig. 3. Sample tracking results of the evaluated algorithms on twelve challenging image sequences. The figures are arranged in the same order as Table. 1.

Occlusion. We test several sequences (*Car6*, *Caviar2* and *Woman*) with severe or long-term partial occlusions. In the *Car6* sequence, only the ℓ_1 , TLD and the proposed methods are able to track the object when the long-term occlusion happens (e.g., *Car6* #528). Note that the ℓ_1 tracker involves occlusion resolving scheme, and the TLD method employs a detector to reacquire the object. The *Caviar2* and *Woman* sequences contain scale change, partial occlusion and interference of similar objects. Most of the trackers lock onto a wrong object after occlusion (e.g., *Caviar2* #240 and *Woman* #130). In contrast, our method achieves stable performance in the entire sequence.

Illumination Change. The tracked objects in the *David*, *Shaking* and *Singer2* sequences undergo significant illumination changes and pose variations. The Frag, IVT, L1, TLD and MIL methods can not handle the appearance variations caused by illumination changes together with pose variations (e.g., *David* #178 and *Shaking* #62), whereas the VTD and SCM methods perform better. In the *Singer2* sequence, the contrast between the foreground and the background is very low. Our method success to track the object accurately, but most trackers drift away at the beginning of the sequence (e.g., *Singer2* #64).

Other Challenges. We test three sequences where the objects suffer other challenges including in-plane rotation (*Panda*), motion blur (*Jumping*) and background clutters (*Board*). Overall, the SCM and our method perform well whereas the other trackers fail to track the objects.

6 Conclusion

We have presented a robust appearance model for visual tracking via a structural object representation strategy and online metric learning. Experiments on twelve challenging sequences demonstrate the robustness of our tracker compared with seven state-of-the-art tracking methods.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China (NSFC) under grant No. 61203291 and the 973 Program of China under grant No. 2012CB720000.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR, vol. 1, pp. 798–805 (2006)
2. Babenko, B., Yang, M., Belongie, S.: Robust object tracking with online multiple instance learning. PAMI 33(8), 1619–1632 (2011)
3. Isard, M., Blake, A.: Condensation conditional density propagation for visual tracking. IJCV 29(1), 5–28 (1998)
4. Jia, X., Lu, H., Yang, M.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR, pp. 1822–1829 (2012)
5. Jiang, N., Liu, W., Wu, Y.: Learning adaptive metric for robust visual tracking. TIP 20(8), 2288–2300 (2011)
6. Kalal, Z., Matas, J., Mikolajczyk, K.: Pn learning: Bootstrapping binary classifiers by structural constraints. In: CVPR, pp. 49–56 (2010)
7. Kwon, J., Lee, K.: Visual tracking decomposition. In: CVPR, pp. 1269–1276 (2010)
8. Li, P., Hastie, T., Church, K.: Very sparse random projections. In: SIGKDD, pp. 287–296 (2006)
9. Li, X., Shen, C., Shi, Q., Dick, A., van den Hengel, A.: Non-sparse linear representations for visual tracking with online reservoir metric learning. In: CVPR, pp. 1760–1767 (2012)
10. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: CVPR, pp. 1313–1320 (2011)
11. Liu, L., Fieguth, P.: Texture classification from random features. TPAMI 34(3), 574–586 (2012)
12. Mei, X., Ling, H.: Robust visual tracking using ℓ_1 minimization. In: ICCV (2009)
13. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV 77(1), 125–141 (2008)
14. Shalev-Shwartz, S., Singer, Y., Ng, A.: Online and batch learning of pseudo-metrics. In: ICML (2004)
15. Wang, D., Lu, H., Yang, M.: Online object tracking with sparse prototypes. TIP 22(1), 314–325 (2012)
16. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR, pp. 3360–3367 (2010)
17. Wang, X., Hua, G., Han, T.X.: Discriminative tracking by metric learning. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 200–214. Springer, Heidelberg (2010)
18. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801 (2009)
19. Zhong, W., Lu, H., Yang, M.: Robust object tracking via sparsity-based collaborative model. In: CVPR, pp. 1838–1845 (2012)

Polygon-Location Method Based on Uyghur Text Regional Rules

Halidan Abudureyimu^{1,*}, Renren Deng¹, Kuerban Maitimusha², and Nana Yang¹

¹ Department of Electrical Engineering, XinJiang University, Urumqi, China

² Department of Mechanical Engineering, XinJiang University, Urumqi, China

halidana@xju.edu.cn

Abstract. Uyghur location for tilted text is of great significance for character recognition. Existing text region localization method is not effective for Uyghur text region. This paper proposed a new method to solve this problem. Firstly extracting rule feature of candidate regions, striking out the noise, then correcting the angle of the tilted text to detect the non-text area; Finally proposing polygon-location algorithm based on the special writing style of Uyghur, which can reduce the interference coming from the background in the process of recognition. The location accuracy is more than 91%.

Keywords: rule feature, angle-correction, polygon-location.

1 Introduction

With the rapid development of modern multimedia and network information technology, a large number of video images appear in the digital library, TV broadcasts and the internet. Considering the similarities with Uyghur and Arabic, the study of the Uyghur character recognition [1] in the video image, which is conducive to improve the level of information processing in Minority, also provides a reference for the surrounding Arabic countries, so as to promote the communication between Xinjiang and neighboring countries.

General text region location algorithms are based on the texture and edge information [2][3]. Image texture can be achieved when using wavelet transform [4] and means algorithm [5] combined with morphological processing, and the corresponding color strokes characteristics can be achieved by BP neural network [6] training. We can set some rules for complex multi-line text, such as the width, height, area [7], make angle correction [8], and further narrow the possible candidate text area. For the extracted candidate region, compare the angle, height, center of gravity of sub-area, merge similar region by using optimal properties [9] and opening and closing operation method; then establish the decision tree and classify the text and non-text area according to distance between the characters, angle, linear feature and finally extract the expected bounding rectangle of text area. But these methods are not suitable for Uyghur text image.

* Research Directions: pattern recognition, Video Image Processing ; Grant from The National Natural Science Funds (61163026、60865001).

2 Algorithm Descriptions

2.1 Extraction of Candidate Region

The obvious difference between text and background are edge features, but some background noise is similar with the edge of the text. As shown in Fig.1(a), firstly using Sobel operator to extract edge information. Then we got Fig.1(b) by morphological expansion, the positioning bounding rectangle is shown in Fig.1(c), which contains many connected area. We label every part from left to right in original picture, the pixel values of each part is equal to the ranked number. Finally, obtain the coordinate of the rectangle, rectangular positioning for each part is applied:

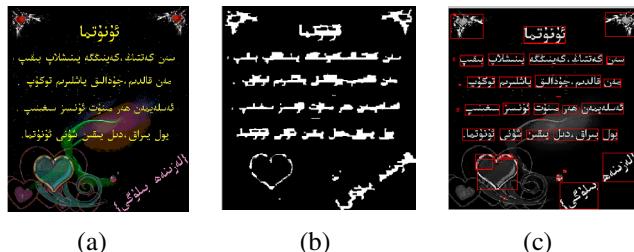


Fig. 1. Candidate Regions (a) Original image (b) Morphological processing (c) location

$$S_{\text{mark}} = \{1 \ 2 \ 3 \ 4 \ 5 \ \dots R-1 \ R\} . \quad (1)$$

$$S_b = \{b_1 \ b_2 \ b_3 \ b_4 \ b_5 \ \dots b_{R-1} \ b_R\} . \quad (2)$$

$$b_r = (x \ y \ \text{width} \ \text{height}) \quad (3)$$

S_{mark} is the set of pixel grayvalue; R is the number of the candidate area; S_b is the set of positioning rectangle; b_r represents the attributes: (x, y) is the left vertex coordinate of bounding rectangle; width, height respectively represent the width and height.

2.2 Determination of Candidate Region

After getting lots of candidate regions, the next step is to determine whether the region is the text area. We set the text width, aspect ratio, area of the rectangle of Uyghur text (in general pictures) in formula (4) to (6).

$$\text{width} > 15 \quad (4)$$

$$\text{AspectRatioR} = \text{Height} / \text{width} < 2 \quad (5)$$

$$\text{Area}_{b_r} = \text{width} * \text{height} > 400 \quad (6)$$

Candidate region whose rectangle area is less than 400, width is less than 15 and aspect ratio is more than 2 is disregarded. Small pots and segment are disregarded as well, finally we got Fig.2(a); sum of pixels in each unicrom region recorded as

Area_{br} . Determine whether it is a text region by ratio of pixels sum in rectangle area. The pixel value of each labeled area is normalized by formula (7):

$$S_R(i, j) = \begin{cases} 1 & \text{if } S_R(i, j) \geq 1 \quad (y < i < y + \text{height}, x < j < x + \text{width}) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$\text{Area}_R = \sum_{i=x}^{x+\text{width}} \sum_{j=y}^{y+\text{height}} S_R(i, j) \quad (8)$$

$$\text{occupancyratio}_R = \frac{\text{Area}_R}{\text{Area}_{\text{br}}} \leq \frac{1}{8} \quad (9)$$

Area_R represents area of connected region labeled as R. We set the ratio of R region and its positioning rectangle in formula (9). Some unfit region whose ratio less than 1/8 is deleted as shown in Fig.2 (b). The benefit of this approach is being able to retain the smaller text area, and delete the larger noise and angle correction could remove the noise with similar features of text region, and the result is shown in Fig.2(c).

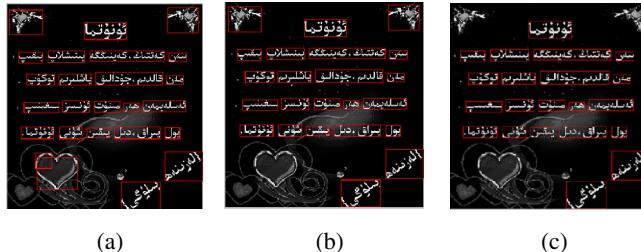


Fig. 2. Results of step by step candidate region determination (a) First candidate region (b) Second candidate region (c) Third candidate region



Fig. 3. Angle correction of non-text



Fig. 4. Angle correction of text

Fig.3 and Fig.4 are respectively a non-text area and text area from Fig.2(b), which shows the process of angle correction. The candidate area rotated by different angles, record the height value in the case of the location rectangle has the smallest area. The height value of non-text correction is still high, however the text correction is similar with formal height, we set the threshold value of the height:

$$\Delta Area_{b_R}(\theta) = Area_{b_R}(\theta + 5) - Area_{b_R}(\theta) \quad (-45 \leq \theta \leq 40) \quad (10)$$

$$Area = [Area_{b_R}(\theta - 45), Area_{b_R}(\theta - 45) \cdots Area_{b_R}(\theta) \cdots Area_{b_R}(\theta + 40)] \quad (11)$$

$$Area_{opt\theta} = \min(Area) \quad (12)$$

$$height_{minarea} < 45 \quad (13)$$

The above formulas (10) to (13) show an area comparison between different angles. The minimum area corresponds to the angle θ , the height is $height_{minarea}$. The algorithm deletes the height of candidate area more than 45, finally obtained the positioning in Fig.2(c).

2.3 Determination of Large Font Text Area

The Uighur text writing bumps ups and downs. There might be many texts with different sizes. As shown in Fig.5 (a), these title texts are generally very important information. So in order to retain such information, separate larger candidate region of the real text from non-text, we have improved the algorithm after angle correction:

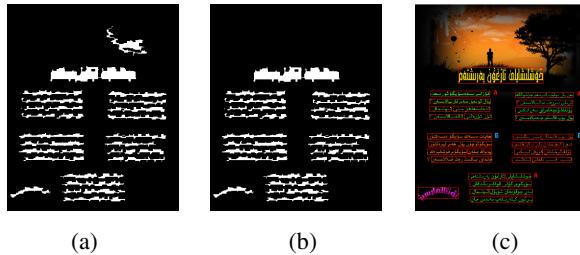


Fig. 5. Determination of regions with large font text (a) Judged region (b) Judged result (c) location

$$s_R(i, j) = \begin{cases} R & \frac{width}{height} \geq 3 \text{ and } height > 45 \\ 0 & \frac{width}{height} < 3 \text{ and } height > 45 \end{cases} \quad (14)$$

Formula (14) shows that if the height of the rectangle is greater than 45, the width to height ratio is greater than 3, it is regarded as large font text, not greater than 3 as the noise. Fig.5(b) is the result after removing the noise region. Fig.5(c) is the positioning result.

3 Polygon Positioning

3.1 Principle of Polygon Positioning

The rectangle rules are easy to be described and more mature in the application, so the selection of the candidate region starting from the rules of the rectangle. But in the

end we chose polygon positioning algorithm, which compared to the rectangular positioning has made improvements on the complexity and accuracy of the algorithm. As shown in Fig.8, the First, start from the rightmost point $I_0(i_0, j_0)$ of the candidate region, the point I_0 is of the center, select the point I which has the smallest counterclockwise rotating angle in the rest of the non-zero pixels, and then from the new center I, repeat the previous method to find another center point, continue until a new point overlapping with the point I_0 .

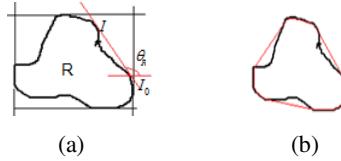


Fig. 6. Polygon-location (a) Gain vertices (b) Location result

The algorithm is as the formula (15):

$$\theta_R = \begin{cases} 2\pi - \arctan \frac{|j_0 - j_1|}{|i_0 - i_1|} & \left(\begin{array}{l} i_4 \geq i_0 \geq i_3, j_4 \geq j_0 \geq j_3 \\ i_1 \geq i \geq i_0, j_0 \geq j \geq j_1 \end{array} \right) \\ 2\pi + \arctan \frac{|j_0 - j_1|}{|i_0 - i_1|} & \left(\begin{array}{l} i_1 \geq i_0 \geq i_4, j_4 \geq j_0 \geq j_1 \\ i_1 \geq i \geq i_0, j_0 \geq j \geq j_1 \end{array} \right) \\ \pi - \arctan \frac{|j_0 - j_1|}{|i_0 - i_1|} & \left(\begin{array}{l} i_1 \geq i_0 \geq i_2, j_1 \geq j_0 \geq j_2 \\ i_0 \geq i \geq i_2, j_0 \geq j \geq j_2 \end{array} \right) \\ \pi + \arctan \frac{|j_0 - j_1|}{|i_0 - i_1|} & \left(\begin{array}{l} i_2 \geq i_0 \geq i_3, j_3 \geq j_0 \geq j_2 \\ i_0 \geq i \geq i_3, j_3 \geq j \geq j_0 \end{array} \right) \end{cases} \quad (15)$$

We set the uppermost point, the lowermost point, the very left and very right point respectively as $(i_2, j_2), (i_4, j_4), (i_3, j_3), (i_1, j_1)$ select the minimum θ_R which is corresponding to the center point. In accordance with the formula (17), the next center point is calculated until it overlaps with the starting center, then link all the vertex to be polygon.

3.2 Polygon Positioning of Uighur Text

Uyghur writing style is special with the characteristic of notable horizontal base line. As shown in Fig.7, the strokes of the text are extending up and down from the baseline, the overall shape of the text shows no rules.

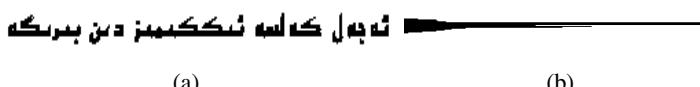


Fig. 7. Uyghur text base line (a) Uyghur font (b) Horizontal projection

Fig.8 is polygon positioning. For Chinese text region, polygon positioning is almost a rectangle, but for Uyghur this method can reduce the background area in the process of Uyghur positioning.



Fig. 8. Polygon-location

4 Experimental Results and Analysis

400 pictures were selected for experiment, which have various forms of fonts, including horizontal, tilted, arched arranged. Some of the pictures are horizontal text, some are horizontal text mixed with tilt fonts or arched ones, fully reflects the comprehensiveness of the polygon location method. Table 1 shows the statistics of positioning results.

Table 1. Statistics of location

Quantity	Location	Accuracy	Incomplete	wrong
400	367	91.75%	4.75%	3.5%

Experiments show that the polygon location algorithm has outstanding effect for Uyghur text. General text region location method is effective for Chinese and English whose character bounding is rectangle and character is isolated with each other ,but boundary of Uyghur text is irregular, bounding rectangle may contains many noise which will affect the recognition as shown in Fig.9(a).Polygon method can reduce the background noise, this would lead a better subsequently extract process which lead a higher resolution for precise recognition. So it is a novel positioning algorithm.



Fig. 9. Uyghur text region location experiments comparison(a)rectangle bounding (b)polygon-location

5 Conclusion

This paper presents an location algorithm for Uyghur text, which includes four steps: 1) Adopting Sobel operator combined with morphological operators to gain the candidate region; 2) Set the rules of candidate region to reduce the text area; 3) Make angle correction for the candidate region, then confirm the text area; 4) Polygon positioning. Experiments show that the polygon location algorithm performs remarkably for video image containing tilt text. It has actual significance of the research for text recognition.

References

1. Abudureyimu, H., Jume, E., Maitimusha, K., Hao, H.: Uighur Character Recognition Based on Adaptive Models. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 476–479 (2012)
2. Kumar, M., Lee, G.: Automatic Text Location from Complex Natural Scene Images, 594–597. IEEE (2010)
3. Wang, Y., Tanaka, N.: Text String Extraction from Scene Image Based on Edge Feature and Morphology. In: The Eighth IAPR Workshop on Document Analysis Systems, pp. 323–328
4. Leon, M., Vilaplana, V., Gasull, A., et al.: Region-Based Caption Text Extraction. In: 2010 11th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4 (2010)
5. Lee, S., Cho, M.S., Jungz, K., et al.: Scene Text Extraction with Edge Constraint and Text Collinearity. In: 2010 International Conference on Pattern Recognition, pp. 3983–3986 (2010)
6. Li, N.-Y., Liang, Y.-M., Zhang, S., et al.: Text Location in Complex Color Images Based on BP Neural Network. *Acta Photonica Sinica* 38(10), 2713–2715 (2009)
7. Jung, J., Lee, S.H., Cho, M.S., Kim, J.H.: Scene Text Extractor Using Touchscreen Interface. *ETRI Journal* 33(1), 78–88 (2011)
8. Rengreng, D., Halidan, A.: An Adaptive-angle-method For Uyghur Location. In: 2012 Spring World Congress on Engineering and Technology, pp. 301–304 (2012)
9. Alves, W.A.L., Hashimoto, R.F.: Text Regions Extracted from Scene Images by Ultimate Attribute Opening and Decision Tree Classification. In: 2010 23rd SIBGRAPI - Conference on Graphics, Patterns and Images, pp. 360–367 (2010)

Study on the Electromagnetic Performance of Hydroelectric Generator Based on Intelligent Control

Xin Shi¹, Jianhua Lin², Yuxiang Wang¹, and Heping Liu²

¹ Department of Electronic Engineering, Shenyang Aviation Vocational and Technical College, Shenyang, P.R. China, 110034

² School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, P.R. China, 100083

Abstract. The control method applied to our hydraulic generator product is still the traditional PID control. This method can't realize self-tuning of control parameters online. To enhance the adaptability of the system, the intelligent control technology is applied to the hydraulic turbine generator set which has effectively improved the dynamic operation performance in different cases. Through theoretical derivation and simulation analysis, we summarized the control performance of improved particle swarm optimization (PSO) method is superior to the fuzzy PID control, when both of them are applied to the hydroelectric regulating system and excitation control system.

Keywords: Hydroelectric generator, governor, excitation controller, PSO.

1 Introduction

With the expanding of power industry, the market demand of large-sized hydraulic generator product is greater and greater. On the one hand the capacity of hydraulic turbine generator unit is further improved, but on the other the power generated by the hydroelectric generator set is more safe and reliable, also the frequency and voltage can be kept in a special range near the ratings. As rotating electromagnetic device, either the electric or magnetic field of the large-sized water turbine generator is distributed in the way of field potential. The study and test of intelligent control system becomes main factors restricting further research and development of large-sized turbine generator. Therefore, the study on electromagnetic properties of hydroelectric generator based on intelligent control can directly promote the turbine generator to the field of large capacity, high efficiency and well stability.

2 Modeling of Hydro-turbine Generator System

The structure of hydraulic turbine generator set is shown in Fig.1. It is composed of diversion system, speed governor, generator, turbine, excitation devices.

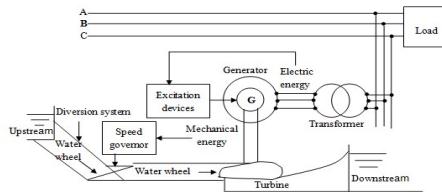


Fig. 1. The basic structure of Hydraulic turbine generator set

The rotary shafts are connected together, when the rotor of the generator that is powered with excitation current, a rotating magnetic field is formed. And the rotating magnetic field cuts across the stator winding generating the three-phase AC power, the balance relationship between generating capacity and power consumption varies constantly. Once the energy balance is broken, the frequency in system will change. In order to stabilize this frequency, the speed regulator needs to be used in the hydraulic turbine generator. The speed governor can detect the speed of the turbine unit, comparing the rotational speed value with the given speed value, then calculating the deviation signal through the control algorithms. The former can use to control the water flow, so as to achieve the balance of water energy and active load. When the reactive power is not balance of the system, the voltage produced by the generator will fluctuate. Thus excitation device need to keep the voltage in a stable threshold, and it can improve the stability of the power angle in the parallel grid generator.

2.1 Model of Hydraulic Generator Set

Turbine generator unit is mainly composed of water diversion system, turbine, generators and load. Its mathematical model can be divided into parallel operation model and single unit operation model. We study the latter of Francis turbine generator. Through Laplace transform, we can achieve the torque and flux formulas[1] as follows:

$$M_t(s) = e_y Y(s) + e_x X(s) + e_h H(s) \quad (1)$$

$$Q(s) = e_{qy} Y(s) + e_{qx} X(s) + e_{qh} H(s) \quad (2)$$

According to Eq.1 and Eq.2, we can draw the structure diagram of transfer function for the hydro-turbine generator, which is shown in Fig.2:

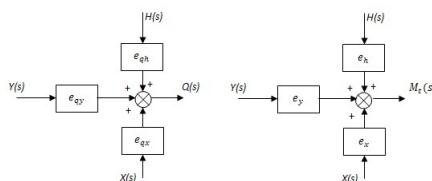


Fig. 2. The block diagram of transfer function for the hydraulic turbine

According to the transfer function[2] of the diversion system and generator load, we can draw the corresponding diagram, as shown in Fig.3:

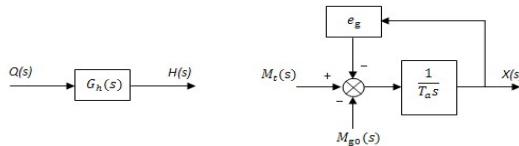


Fig. 3. The block diagram of diversion system and generator load

Normal, the transfer coefficient e_{qx} in Francis turbine can be neglected. So, when using a comprehensive adjusting-parameter e_n to present the self-regulation parameter of the water turbine and the generator load, according to Fig.2 and Fig.3, we can get the structure block diagram of hydroelectric generators set as shown in Fig.4:

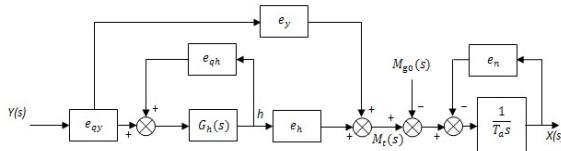


Fig. 4. Block diagram of hydroelectric generators set

2.2 The Model of Excitation System for Generator

The excitation control system is composed of voltage detection, power amplification, synchronous generators and excitation controller, as shown in Fig.5:

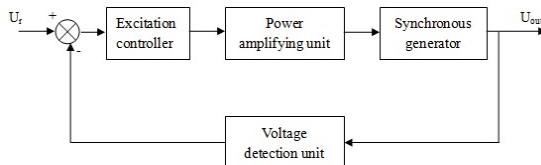


Fig. 5. The block diagram for excitation control system

Excitation controller uses a control algorithm, such as the PID control, fuzzy PID control and some new intelligent control algorithms. The control signal generated by it goes through the power amplifying unit achieving appropriate control power. Excitation controller can directly load the rectified control current to the rotor of synchronous generator, changing magnetic intensity of the rotor by varying excitation current, and thus to control the generator voltage and power factor to keep the

stability of the power system. The voltage detection unit is mainly composed of measuring transformer, rectifier and filter circuit, measurement circuit. It is used to measure the output voltage of the generator and compare the generated signal with the given one. Then the error signal is transmitted to the excitation controller. Usually, the transfer function of the power amplification unit and the voltage detection unit can be described by one order inertial link[3] as follows:

$$G_A(s) = \frac{1}{1 + T_A s} \quad (3)$$

$$G_R(s) = \frac{1}{1 + T_R s} \quad (4)$$

Eq.3 is the transfer function of the power amplification unit, where T_A is the time constant for amplifying circuit link. We give it the value of 0.7 in the model of this paper. Eq.4 is the transfer function of the voltage detecting unit, where T_R is the time constant for the filter circuit. Its value is 0.035 in the model of this paper.

3 A Modification in Particle Swarm Optimization Algorithm

We assume that the set $X = (X_1, X_2, \dots, X_n)$, consist of the particles, flight in a D-dimensional space, where the i^{th} particle is described as a D-dimensional vector $X_i = (X_{i1}, X_{i2}, \dots, X_{iD})^T$, representative of the position of the i^{th} particle in this space. We calculate the fitness value of each particle corresponding to the position X_i according to the fitness function, the i^{th} particle's velocity can be expressed as $V_i = (V_{i1}, V_{i2}, \dots, V_{iD})^T$, its individual extreme can be expressed as $P_i = (P_{i1}, P_{i2}, \dots, P_{iD})^T$, population groups extreme $P_g = (P_{g1}, P_{g2}, \dots, P_{gD})^T$. In each process, the particles may go through the velocity and the position equation which are represented by individual and global extreme to update its speed and position[4], i.e.:

$$V_{id}^{k+1} = \omega V_{id}^k + c_1 r_1 (P_{id}^k - X_{id}^k) + c_2 r_2 (P_{gd}^k - X_{gd}^k) \quad (5)$$

$$X_{id}^{k+1} = X_{id}^k + V_{id}^{k+1} \quad (6)$$

Where ω is the inertia weight and it is the momentum of the particle movement. k is the current iteration number; $d = 1, 2, \dots, D$; $i = 1, 2, \dots, n$, V_{id} accounts for the velocity of the particle; r_1 and r_2 are random numbers in the range [0,1]; c_1 and c_2 are positive constants in standard PSO algorithm, called acceleration coefficients. In order to avoid particle being blind search in the space, the speed and position of the particle are confined between $[-V_{\max}, V_{\max}]$ and $[-X_{\max}, X_{\max}]$.

When using standard particle swarm optimization algorithm, the inertia weight usually represented by a decreasing linear equation[5]:

$$\omega = (\omega_{ini} - \omega_{end})(T - t)/T + \omega_{end} \quad (7)$$

Where, ω_{ini} is the initial inertia weight of 0.9, ω_{end} accounts for the terminal inertia weight of 0.4; let T be the maximum iteration number, t be the current one.

We combine hydro-generator system with the characteristics of particle swarm optimization, using the law that inertia weight and two learning factors to improve the optimization for standard particle swarm at the same time. We form the IPSO (Improved PSO) algorithm. The influence that the current iteration speed can be controlled by inertia weights, the bigger inertia weights is, the bigger PSO's searching ability for the whole is, and the smaller it is, the bigger its ability for the partial. As Eq.5, when learning factor c_1 is greater than c_2 , the particles move closer to their own optimal position. To the contrary, the particles move closer to the global optimal position. Therefore, this method allows particles to have good global search ability in the search, to effectively improve the particle premature convergence defects. In this paper, we use the expression of the cosine function to improve the learning factor, this is different from other literatures which using the improved method of linear decreasing or increasing. This expression in the vicinity of the boundary value of the learning factor is more gently than the linear method, the expression is defined as:

$$\begin{cases} c_1 = \varphi_1 + 0.5 \cos[(t/T)\pi] \\ c_2 = \varphi_1 - 0.5 \cos[(t/T)\pi] \end{cases} \quad (8)$$

Where, φ_1 and φ_2 are constant, and set $\varphi_1 = \varphi_2 = 2$; let T be the maximum iteration number and t be the current one. In Eq.8, set $c_1 = c_2$, $\varphi_1 = \varphi_2 = 2$, then calculate $t/T = 1/2$. When half of the iteration numbers, the influences of the two learning factors are equal, this can meet the changing needs of the learning factor.

4 Simulation Study

In order to verify the advantages of particle swarm optimization in hydro-generator system, this section is simulated and selected the turbine parameters as follow:

We set temporary droop $Bt=0.8$, transient feedback time constant $Td=3.36$, servomotor time constant $Ty=0.2$, flow-accelerating time constant $Tw=1.0$, transfer coefficient $Ey=1.0$, $Eh=1.5$, $Eqy=1.0$, $Eqh=0.5$, Inertial unit time constant $Ta=5$, Unit self-balancing factor $En=1.0$. Hydro-generator set startup process frequency response curve shown as Fig.6, the improved particle swarm algorithm optimizes PID control works best. Fig.7 is a characteristic diagram, under stable operating conditions of the unit, adding 40Hz interference frequency after. We can see the interference effect of the controller designed by intelligent control algorithm is better than the traditional PID controller, and the interference effect of the controller designed by improved particle swarm optimization algorithm is better than Fuzzy PID controller.

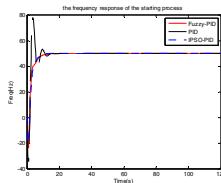


Fig. 6. The comparison of the set frequency response of the starting progress

PID controller optimized by improved particle swarm optimization also has good characteristics of synchronous generator excitation control. As shown in Fig.8.

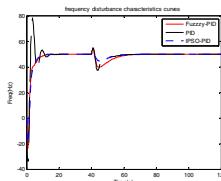


Fig. 7. The comparison of the set frequency disturbance characteristics curves

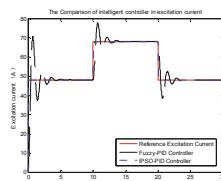


Fig. 8. The follow characteristic curve of controller for a given excitation values

5 Conclusion

This paper focuses on the research of hydro-generator speed control system and excitation control system, we make simulation on hydro-generator set governor and excitation regulator using traditional PID control method , fuzzy PID control method and IPSO algorithm, PID optimization control method, and fully verify the superiority of the intelligent control algorithm, and also proved that the proposed improved PSO algorithm has better performance characteristics than the fuzzy algorithm.

References

1. Liao, Z., Li, Z.: The Fuzzy Control Simulation of the Hydro-Turbine Governing system. Journal of Nanchang College of Water Conservancy and Hydroelectric 4, 17–20 (2001)
2. Li, P., Cai, W., Xiao, Z.: Fuzzy-PID controller of hydro-turbine regulating system and its simulation model. Northwest Water Power 12(3), 88–90 (2004)

3. Han, Y., Xie, X., Cui, W.: Status Quo and Future Trend in Research on Synchronous Generator Excitation Control. *J. Tsinghua Univ.* 41(4/5), 142–146 (2001)
4. Wang, C., Duan, X., Liu, X.: A Modified Basic Particle Swarm Optimization Algorithm. *Computer Engineering* 30(21), 435–439 (2004)
5. Li, J., Sun, X., Li, S., Li, R.: Improved Particle Swarm Optimization Based on Genetic Hybrid Genes. *Computer Engineering* 2, 1021–1025 (2008)

Text-Independent Phoneme Segmentation via Learning Critical Acoustic Change Points

Peng Teng, Xiabi Liu,, and Yunde Jia

School of Computer, Beijing Institute of Technology, China
Beijing Laboratory of Intelligent Information Technology, China
tengpeng,liuxiabi,jiayunde}@bit.edu.cn

Abstract. The conventional methods of automatic text-independent phoneme segmentation detect phoneme boundaries via calculating the acoustic changes along speech signals followed by a peak picking procedure according to user-defined rules. Instead, this paper presents a learning-based method in which the phoneme boundaries are viewed as critical points in the acoustic change context of speech signals. First, we adopt a metric learning procedure in the calculation of acoustic changes, in order to make the acoustic changes at phoneme boundaries more discriminative. Then, latent-dynamic conditional random field is used to model the acoustic change context of speech signals for the detection of phoneme boundaries. The experiments demonstrate that our method outperforms the rule-based methods reported in previous work.

Keywords: Text-Independent Phoneme Segmentation, Latent-Dynamic Conditional Random Field, Metric Learning.

1 Introduction

Text-independent phoneme segmentation is the partitioning of a continuous speech signal into discrete and non-overlapping units without the text content of the speech. It is a fundamental element of phoneme-based speech recognition system which suggests a candidate framework for developing the next generation speech processing techniques [1], and it is also used for the automatic phonemic analysis of large amounts of speech data. Unlike written language, a speech signal does not contain explicit markers to indicate phoneme boundaries, and it is difficult to define what a boundary is like in its acoustic signal. In view of the important uses and the challenges in the implementation, text-independent phoneme segmentation has received much research interest in the last decade [2–10].

Most of the previous text-independent phoneme segmentation methods [2, 3, 8, 10] focus on detecting salient change points in speech signals. These methods pick salient points from the acoustic changes along speech signals according to user-defined rules, and take the points as hypothesized boundaries of adjacent phoneme segments. The major drawback of these methods is that finding the proper threshold values with which users describe the rules can be difficult.

Some recent methods [4, 6, 7] view the phoneme segmentation problem as an optimal segmentation problem over a sequential data. They adopt clustering-based algorithms to obtain frame clusters along speech signals, and take each cluster as a phoneme segment. However, these algorithms usually require the number of clusters (i.e., the number of phoneme segments) as their inputs, which makes these methods not fully text-independent.

In this paper, we propose a novel method for text-independent phoneme segmentation via learning critical acoustic change points as the phoneme boundaries. Our motivation is to achieve the automatic segmentation by simulating how a young infant does, since the situation is exactly the circumstances under which infants have to learn to speak and understand their native language [5]. Two important discoveries in psycholinguistics have informed us about the nature of the innate skills that infants bring to the phoneme segmentation task [11]. The first is called categorical perception which focuses on the discrimination of the acoustic events that distinguish phonemes. The second is called categorization which refers to the phenomenon that infants can group together different sounds that they clearly hear as distinct. Inspired by the above discoveries, we view phoneme boundaries as critical points in acoustic changes along speech signals. We use latent-dynamic conditional random fields [12] to model the context of these points in acoustic changes. In the measurement of the acoustic changes, a metric learning procedure [13] is employed to make the differences among frames from different phoneme classes relatively larger than those from the same class, enhancing the salience of the changes at the boundaries. Experiments demonstrate that the segmentation quality of our method is better than those of previous rule-based text-independent phoneme segmentation methods reported on TIMIT [14] database.

2 Speech Segmentation Modeling Using Graphic Models

In a graphical model with a graph G , a vertex of G denotes a random variable and an edge denotes the dependence between two variables. A speech signal which consists of T time-frames and represents a phoneme sequence $\mathbf{Ph} = \{Ph_1, \dots, Ph_M\}$ can be modeled with a graphical model shown in Fig. 1(a), where $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_T]$ is an observed L -dimension parameter vector sequence of the T time-frame signal, $\mathbf{R} = [R_1, \dots, R_T]$ is a phoneme sequence corresponding to each time frame, $R_t \in \mathbf{Ph}$, and S_t is a sub-state of the phoneme R_t . Let \mathcal{S}_{R_t} be the set of possible sub-states for phoneme R_t , and \mathcal{S} denote the set of all possible sub-states for all possible phonemes, $S_t \in \mathcal{S}_{R_t} \subseteq \mathcal{S}$. From the speech model, we derive a phoneme boundary model by considering the changes among adjacent time frames, as shown in Fig. 1(b), where $B_t = (R_t \neq R_{t+1}) \in \{1, 0\}$ is a boundary label indicating any phoneme boundary presence or not at the t -th time frame, \mathbf{y}_t is a feature vector representing the acoustic change around the t -th time frame, and H_t is the change state on set \mathcal{S} depending on the boundary label B_t . Let \mathcal{H}_{B_t} denote the set of possible change states associated with B_t and \mathcal{H} denote the set of all possible change states, $H_t \in \mathcal{H}_{B_t} \subseteq \mathcal{H}$. The variables in

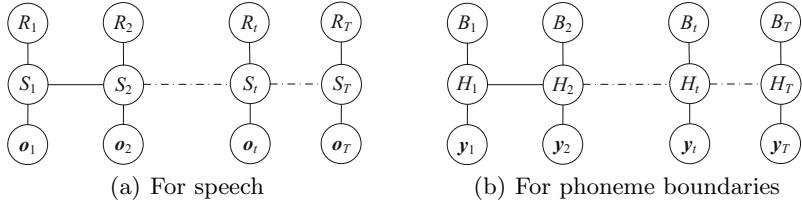


Fig. 1. Graphical models representation of LDCRF for phoneme segmentation

$H = [H_1, \dots, H_T]$ are difficult to be directly observed in speech signals and will therefore form a set of *hidden states* in the model. Under the above definitions, the phoneme segmentation of a speech signal is to predict a boundary label sequence $\mathbf{B} = [B_1, \dots, B_T]$ along the acoustic change sequence $\mathbf{Y} = [y_1, \dots, y_T]$ of the signal.

In our method, the mapping between acoustic change sequence \mathbf{Y} and phoneme boundary label sequence \mathbf{B} is learned by using latent-dynamic conditional random fields (LDCRF). In LDCRF with the structure depicted in Fig. 1(b), the latent conditional model is defined as

$$p_{\theta}(\mathbf{B}|\mathbf{Y}) = \sum_{\mathbf{H}} p_{\theta}(\mathbf{B}|\mathbf{H}, \mathbf{Y}) p_{\theta}(\mathbf{H}|\mathbf{Y}), \quad (1)$$

where θ is the parameter set of the model. For practicability, the model is restricted to have disjoint sets of hidden states associated with each boundary label. Therefore, hidden state sequences which have any $H_j \notin \mathcal{H}_{B_j}$ will have $p_\theta(\mathbf{B}|\mathbf{H}, \mathbf{Y}) = 0$ by definition, and Eq. (1) can be rewritten as

$$p_{\theta}(\mathbf{B}|\mathbf{Y}) = \sum_{\mathbf{H}: \forall H_j \in \mathcal{H}_{B_j}} p_{\theta}(\mathbf{H}|\mathbf{Y}). \quad (2)$$

$p_{\theta}(\mathbf{H}|\mathbf{Y})$ is defined by using the usual conditional random field formulation [15]:

$$p_{\boldsymbol{\theta}}(\mathbf{H}|\mathbf{Y}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{Y})} \exp\left(\sum_{k=1}^K \theta_k \cdot \mathbf{F}_k(\mathbf{H}, \mathbf{Y})\right), \quad (3)$$

where the partition function Z is defined as

$$Z_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{\mathbf{H}} \exp \left(\sum_{k=1}^K \theta_k \cdot \mathbf{F}_k(\mathbf{H}, \mathbf{Y}) \right). \quad (4)$$

\mathbf{F}_k is defined as

$$\mathbf{F}_k(\mathbf{H}, \mathbf{Y}) = \sum_{t=1}^T f_k(H_{t-1}, H_t, \mathbf{Y}, t), \quad (5)$$

where each feature $f_k(H_{t-1}, H_t, \mathbf{Y}, t)$ is either a state function $h_k(H_t, \mathbf{y}_t)$ or a transition function $t_k(H_{t-1}, H_t)$, and K is the total number of the feature

functions. $|\mathcal{H}| \times |\mathcal{H}|$ transition functions t_k are defined one for each hidden state pair (H', H'') as

$$t_k(H_{t-1}, H_t) = \begin{cases} 1 & \text{if } H_{t-1} = H' \text{ and } H_t = H'' \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

Each state function is defined one for each “hidden state and feature index” pair (H', l) as

$$h_k(H_t, \mathbf{y}_t) = \begin{cases} \mathbf{y}_t^{(l)} & \text{if } H_t = H' \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

so that the number of the state functions is equal to the length L of the feature vector \mathbf{y}_t times the number of possible hidden states $|\mathcal{H}|$. Therefore, $K = |\mathcal{H}| \times |\mathcal{H}| + |\mathcal{H}| \times L$. Given a training set consisting of n labeled sequences $(\mathbf{B}_i, \mathbf{Y}_i)$ for $i = 1, \dots, n$, the optimal parameter values $\boldsymbol{\theta}^*$ is estimated as

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \left(\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\mathbf{B} | \mathbf{Y}) - \frac{1}{2\sigma^2} \|\boldsymbol{\theta}\|^2 \right) \quad (8)$$

using the algorithm mentioned in [12]. Denote $p_{\boldsymbol{\theta}^*}(\mathbf{B} | \mathbf{Y})$ by $p(\mathbf{B} | \mathbf{Y})$ for simplicity. Given the acoustic change sequence \mathbf{Y} of an unseen speech signal, its phoneme boundary label sequence \mathbf{B} is estimated by maximizing the conditional model:

$$\mathbf{B}^* = \arg \max_{\mathbf{B}} p(\mathbf{B} | \mathbf{Y}). \quad (9)$$

Again, assuming each boundary label is associated with a disjoint set of hidden states, Eq. (9) can be rewritten as

$$\mathbf{B}^* = \arg \max_{\mathbf{B}} \sum_{\mathbf{H}: \forall H_t \in \mathcal{H}_{B_t}} p(\mathbf{H} | \mathbf{Y}). \quad (10)$$

When estimating the boundary label of the t -th time frame, the marginal probabilities $p(H_t = a | \mathbf{Y})$ are computed for all possible hidden states $a \in \mathcal{H}$. Then, the marginal probabilities are summed according to the disjoint sets of hidden states \mathcal{H}_{B_t} , and the label associated with the optimal set is chosen. In our context, we focus on the the summed marginal probabilities when $\mathbf{B}_t = 1$, i.e., $p(\mathbf{B}_t = 1 | \mathbf{Y})$. We assign $\mathbf{B}_t = 1$, if $p(\mathbf{B}_t = 1 | \mathbf{Y})$ is greater than a given threshold η . A further post-processing on the estimated \mathbf{B} is required: for a sequence of consecutive time frames $[B_{t_1}, \dots, B_{t_2}] = [1, \dots, 1]$, check its corresponding marginal probability sequence $[p(B_{t_1} = 1 | \mathbf{Y}), \dots, p(B_{t_2} = 1 | \mathbf{Y})]$, and only take the frames with peak values as hypothesized boundaries.

3 Learning Discriminative Measurement for Acoustic Changes

In this section we discuss how to measure the acoustic changes derived from observed parameter vector sequence \mathbf{O} , in order to make acoustic changes at

phoneme boundaries more different from those within a phoneme. Most methods of measuring acoustic changes are based on the calculation of distance between two adjacent frames on their parameter vectors. These methods view each dimension of the parameter vectors non-distinctive. That is, if denoting the distance between two observed parameter vectors \mathbf{o}_i and \mathbf{o}_j as

$$D_{\mathbf{W}}(\mathbf{o}_i, \mathbf{o}_j) = \| \mathbf{W}(\mathbf{o}_i - \mathbf{o}_j) \|_2^2 = (\mathbf{o}_i - \mathbf{o}_j)^T \mathbf{W}^T \mathbf{W} (\mathbf{o}_i - \mathbf{o}_j), \quad (11)$$

they set \mathbf{W} to the identity matrix. Instead, we aim at finding a more general \mathbf{W} under which the distance $D_{\mathbf{W}}$ is larger when \mathbf{o}_i and \mathbf{o}_j are from different phoneme classes and smaller when they are from the same phoneme class. Let Ω_c denote the set of samples in the c -th phoneme class. According to linear discriminant analysis (LDA), such \mathbf{W} maximizes the amount of between-class variance relative to the amount of within-class variance. These variances are computed from the between-class and within-class covariance matrices, defined by

$$\begin{aligned} \mathbf{C}_b &= \frac{1}{C} \sum_{c=1}^C (\mu_c - \mu)(\mu_c - \mu)^T \\ \mathbf{C}_w &= \frac{1}{N_C} \sum_{c=1}^C \sum_{\mathbf{o}_i \in \Omega_c} (\mathbf{o}_i - \mu_c)(\mathbf{o}_i - \mu_c)^T \end{aligned}, \quad (12)$$

where C is the number of phoneme classes, μ_c is the mean vector of samples in Ω_c , μ is the global mean vector of all the phoneme class, and the total number of samples is N_C . The expected \mathbf{W} can be formulated and estimated as

$$\begin{aligned} \mathbf{W}^* &= \arg \max_{\mathbf{W}} \text{Trace} \left(\frac{\mathbf{W}^T \mathbf{C}_b \mathbf{W}}{\mathbf{W}^T \mathbf{C}_w \mathbf{W}} \right). \\ \text{subject to : } &\mathbf{W} \mathbf{W}^T = \mathbf{I} \end{aligned} \quad (13)$$

Our measurement of acoustic changes will therefore be derived from $\tilde{\mathbf{O}} = \mathbf{W}^* \mathbf{O}$. To obtain smooth measurement values, we compute the squared l_2 norm of the regression coefficients of $\tilde{\mathbf{O}}$ at t -th time frame as

$$A_t(\tilde{\mathbf{O}}) = \left\| \frac{\sum_{\gamma=1}^{\Gamma} \gamma \cdot (\tilde{o}_{t+\gamma} - \tilde{o}_{t-\gamma})}{2 \sum_{\gamma=1}^{\Gamma} \gamma^2} \right\|_2^2, \quad (14)$$

where Γ represents the number of frames used to compute these regression coefficients. Then, $A_t(\tilde{\mathbf{O}})$ can serve as a discriminative measurement of acoustic change at t -th time frame. Note that, if MFCCs are adopted as \mathbf{O} and \mathbf{W}^* is set to the identity matrix, $A_t(\tilde{\mathbf{O}})$ is equal to the spectral transition measure (STM) which is the measurement of acoustic change used in Dusan's approach [3].

4 Experiments

The evaluation of our method is carried out on the full train set of the TIMIT database in order to be comparable with many previous methods. Correspondingly, we use the full Test set to serve as the training corpus of the models. All

the speech signals are recorded at 16 kHz. We enframe each of the signals using a sliding window of length 32 ms and the window shift of 10 ms. The phoneme associated with a frame is assigned with the one that the most sampling points are labeled with in the transcription. A phoneme boundary may appear in three joint frames, and we only label the middle frame of the three as a phoneme boundary. From each frame, 10-MFCCs and their delta are extracted and adopted as \mathbf{o}_t . Besides, an energy coefficient e_t is also extracted. Therefore, two parameter sequences are obtained for each sentence, i.e., a spectral parameter vector sequence $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_T]$ and an energy parameter sequence $\mathbf{E} = [e_1, \dots, e_T]$.

In the learning of \mathbf{W} , the number of phoneme classes should be determined firstly. The number of phonetic symbols contained in TIMIT is 61. We reduced it to 6 by gathering phonemes with the same manners, i.e., Stop: {b, d, g, p, t, k, dx}, Affricates&Fricatives: {jh, ch, s, sh, z, zh, f, th, v, dh}, Nasals: {m, n, ng, em, en, eng, nx}, Semivowels&Glides: {l, r, w, y, hh, hv, el}, Vowels: {iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h} and Pause: {epi, q, bcl, dcl, gcl, kcl, pcl, tcl, pau, h#}. Phonemes with the same manner sound comparatively similar, and are rarely spoken jointly. Then, we gather all the frames from TIMIT test set and their associated phoneme class labels to estimate \mathbf{W} according to Eq.(13). The feature vectors representing the acoustic changes in LDCRFs contain two parts. The first part is the measurement of spectral change $A_t(\mathbf{WO})$. The second part is the measurement of short-term energy change derived from the regression coefficients of \mathbf{E} , and computed as

$$A_t(\mathbf{E}) = \left\| \frac{\sum_{\gamma=1}^{\Gamma} \gamma \cdot (e_{t+\gamma} - e_{t-\gamma})}{2 \sum_{\gamma=1}^{\Gamma} \gamma^2} \right\|_2. \quad (15)$$

So, the feature vector representing the acoustic changes at t -th time frame is $\mathbf{y}_t = [A_t(\mathbf{E})^\top, A_t(\mathbf{WO})^\top]^\top$ (with the parameter $\Gamma = 2$).

Each sentence in the TIMIT test set is transformed into a feature vector sequence, and its phonetic transcription is transformed into a boundary label sequence, correspondingly, serving as training pairs for LDCRF. We adopt the implementation of LDCRF at <http://sourceforge.net/projects/hcrcf/>, with parameters $|\mathcal{H}_{B_t}| = 3$ and $\sigma = 10$.

The accuracy scores introduced in [10] with a tolerance window of 20 ms are used as standard measurements of the segmentation quality, i.e., Hit Rate (HR), False Alarm Rate (FA), Over Segmentation Rate (OS) and two global scores: F_1 -value and R -value. The performances of Dusan's method [3], Qiao's two methods ("A" for [6] and "B" for [7]) and Khanaghā's method [10] are taken as references, since these methods report on the full train set of TIMIT. In addition, two degraded variations of our methods are implemented and evaluated, in order to demonstrate the improvements arising from the use of LDCRF and the metric learning procedure. The two degraded variations adopt the same LDCRF model as our original methods, but different measurements of acoustic change. The first degraded variation adopts STM as the feature of acoustic changes. The second adopts $\mathbf{y}_t = [A_t(\mathbf{E})^\top, A_t(\mathbf{O})^\top]^\top$ as the feature of acoustic changes, i.e., without the metric learning procedure. The ROC curves of our original method

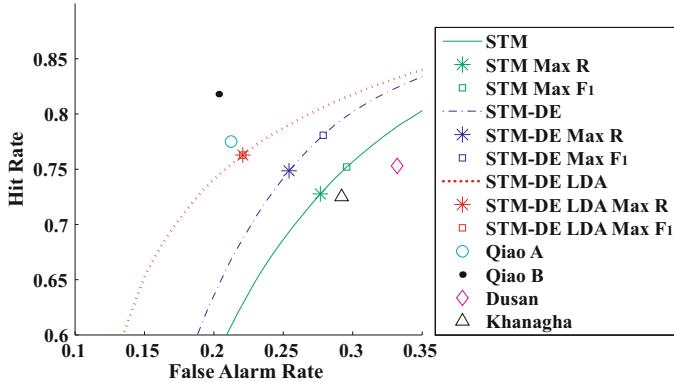


Fig. 2. Comparison of segmentation methods on their ROC curves

and the two degraded variations, as well as the ROC data of the four reference methods, are illustrated in Fig. 2. The two degraded variations and our original method are denoted by “STM”, “STM-DE” and “STM-DE LDA”, respectively. The suffix “Max R” and “Max F₁” denote the “partial” scores when each ROC curve achieves the max global scores, respectively. The details corresponding to Fig. 2 is shown in Table. 1. Using the same measurement of acoustic change, the first degraded variation (STM) outperforms Dusan’s method, which demonstrates the relative advantages of learning-based method to rule-based ones. Our method (STM-DE LDA) outperforms the second degraded variation (STM-DE), and the second degraded variation outperforms the first one (STM). This demonstrates the improvement arising from the metric learning procedure and the energy-derived measurement of acoustic change. The performances of Qiao’s two methods are reported better than those of ours. However, the two methods are both required the number of phoneme segments as their inputs, which makes them not fully text-independent.

Table 1. Detailed scores in the comparison of segmentation methods

Methods Scores		STM-DE LDA	STM-DE	STM	Qiao A	Qiao B	Dusan	Khanagha
HR	(Max R)	0.762	0.747	0.727	0.775	0.818	0.753	0.725
	(Max F ₁)	0.762	0.780	0.751				
FA	(Max R)	0.220	0.253	0.276	0.212	0.206	0.332	0.292
	(Max F ₁)	0.220	0.278	0.295				
OS	(Max R)	-0.023	0.001	0.004	-0.016	0.027	0.128	0.025
	(Max F ₁)	-0.023	0.080	0.066				
R (Max)		0.804	0.784	0.765	0.813	0.834	0.729	0.756
F ₁ (Max)		0.771	0.750	0.727	0.781	0.807	0.708	0.716

5 Conclusions

This paper has presented a learning-based text-independent phoneme segmentation method. We have adopted the metric learning technique in measuring acoustic changes, in order to make the measurement more salient at phoneme boundaries. Then, LDCRF has been used to model the acoustic change context of a speech signal for the detection of phoneme boundaries and the further segmentation. The experiments have demonstrated that our method outperforms the rule-based methods reported on the full Train set of TIMIT database.

Acknowledgments. This work was supported in part by the Natural Science Foundations of China (81171407) and Specialized Research Fund for the Doctoral Program of Higher Education (20121101110035).

References

1. Lee, C.-H., et al.: An overview on automatic speech attribute transcription (ASAT). In: Proc. Interspeech, pp. 1825–1828 (2007)
2. Aversano, G., et al.: A new text-independent method for phoneme segmentation. In: Proc. IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516–519 (2001)
3. Dusan, S., Rabiner, L.: On the relation between maximum spectral transition positions and phone boundaries. In: Proc. InterSpeech, pp. 17–21 (2006)
4. Estevan, Y., et al.: Finding maximum margin segments in speech. In: Proc. ICASSP, vol. 4, pp. IV–937 (2007)
5. Scharenborg, O., et al.: Segmentation of speech: Childs play? In: Proc. International Conference on Spoken Language Processing, pp. 1953–1956 (2007)
6. Qiao, Y., et al.: Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons. In: Proc. ICASSP, pp. 3989–3992 (2008)
7. Qiao, Y., Minematsu, N.: Metric learning for unsupervised phoneme segmentation. In: Proc. Interspeech (2008)
8. Almpandis, G., et al.: Robust detection of phone boundaries using model selection criteria with few observations. IEEE Transactions on Audio, Speech, and Language Processing 17(2), 287–298 (2009)
9. Scharenborg, O., et al.: Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. The Journal of the Acoustical Society of America 127, 1084 (2010)
10. Khanagha, V., et al.: Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism. In: Proc. ICASSP, pp. 4484–4487 (2011)
11. Kuhl, P.K.: Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 5(11), 831–843 (2004)
12. Morency, L.-P., et al.: Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR, pp. 1–8 (2007)
13. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research 10, 207–244 (2009)
14. Garofolo, J.S.: TIMIT: acoustic-phonetic continuous speech corpus. Linguistic Data Consortium (1993)
15. Lafferty, J., et al.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML (2001)

Robustness Analysis of Z-type ZLE Solving

Weibing Li, Wenchao Lao, Xiaotian Yu, Zenghai Chen, and Yunong Zhang

School of Information Science and Technology, Sun Yat-sen University,
Guangzhou 510006, China
zhynong@mail.sysu.edu.cn

Abstract. The general method of Z-type model for online solution of Zhang linear equation (i.e., ZLE or termed time-varying linear equation) is presented. Global exponential convergence of the Z-type ZLE model can be achieved theoretically. In this paper, robustness properties of the Z-type ZLE model in the presence of normal differentiation and dynamics-implementation errors are investigated. Both theoretical analysis and computer-simulation results demonstrate the good robustness of the Z-type model for ZLE solving. In addition, the Z-type ZLE method is compared with G-type ZLE method in the Appendix.

Keywords: Z-type, G-type, Zhang linear equation, robustness, global exponential convergence.

1 Introduction

The online solution of linear equations (including matrix inversion as a closely-related topic) appears to be ubiquitous in science and engineering fields [1–3]. Neural dynamics is now viewed as a powerful alternative to online computation owing to its parallel distributed nature and convenience of hardware implementation [3, 4]. Specifically, neural dynamics is also an efficient method for solving linear equations and related matrix/vector problems [3].

The conventional gradient dynamics (GD) has been substantiated well for static (i.e., time-invariant) problems solving [1, 3]. However, when applied to time-varying problems (i.e., Zhang problems), the G-type model (that is designed based on GD) only works approximately, lagging behind with large solution-errors [3]. In view of this, a special class of neural dynamics, termed Zhang dynamics (ZD), has been proposed for such time-varying problems solving [3–5].

It has been proved that global exponential convergence of the resultant Z-type models (which are designed based on ZD) could be achieved for Zhang linear equation (i.e., ZLE or termed time-varying linear equation) solving [3, 5]. However, in practice, there always exist some realization errors, which is more complicated than the proved ideal situation. In view of this, robustness properties of the Z-type model are investigated in this paper with both normal differentiation and dynamics-implementation errors considered.

2 ZLE Problem Formulation and Z-Type Solver

The problem of ZLE could be generally formulated as [3–5]:

$$A(t)x(t) = b(t), \quad (1)$$

where coefficient matrix $A(t) \in R^{n \times n}$ and vector $b(t) \in R^n$ are smoothly time-varying, while Zhang quotient $x(t) \in R^n$ is the unknown vector to be obtained. If $A(t)$ is nonsingular, the theoretical solution to ZLE (1) is the theoretical Zhang quotient $x^*(t) = A^{-1}(t)b(t)$. To solve ZLE (1), the resultant Z-type model is

$$A(t)\dot{x}(t) = -\dot{A}(t)x(t) - \gamma\mathcal{F}(A(t)x(t) - b(t)) + \dot{b}(t), \quad (2)$$

with design parameter $\gamma > 0 \in R$ and activation-function array $\mathcal{F}(\cdot) : R^n \rightarrow R^n$. Inspired by [3–5], in this paper, the arrays of linear activation function (LAF) and power-sigmoid activation function (PSAF) are exploited.

For detailed comparisons between general Z-type models and G-type models on time-varying problems [i.e., Zhang problems, e.g., ZLE (1)] solving, interested readers can refer to the Appendix of this paper. Regarding the Z-type solution model (2) for solving ZLE (1), we have the following lemma [3, 5].

Lemma 1. *For smoothly time-varying $b(t) \in R^n$ and nonsingular $A(t) \in R^{n \times n}$, if a monotonically-increasing odd activation-function array $\mathcal{F}(\cdot)$ is used, then the state vector $x(t)$ of Z-type ZLE model (2) starting from any initial state $x(0) \in R^n$ globally (exponentially) converges to the time-varying theoretical solution $x^*(t) = A^{-1}(t)b(t)$ which is also termed the theoretical Zhang quotient.*

The above lemma guarantees the excellent convergence of Z-type ZLE model (2) under ideal conditions, e.g., with no errors involved and using LAF or PSAF. However, realization errors always exist in hardware implementation. Thus, in the ensuing sections, robustness properties of Z-type ZLE model (2) are investigated with normal differentiation and dynamics-implementation errors involved.

3 Robustness Analysis of Z-Type ZLE Solver

In the hardware implementation of Z-type ZLE-solving model (2), the differentiation errors about matrix $A(t)$ and/or vector $b(t)$ as well as the dynamics-implementation error (termed collectively as the model-implementation error) appear most frequently. Therefore, let us consider the following dynamic equation which might depict a perturbed Z-type ZLE model:

$$A\dot{x} = -(\dot{A} + \Delta_{m_1})x - \gamma\mathcal{F}(Ax - b) + \dot{b} + \Delta_{m_2}, \quad (3)$$

where $\Delta_{m_1}(t) \in R^{n \times n}$ denotes the differentiation error of matrix $A(t)$, and $\Delta_{m_2}(t) \in R^n$ denotes the dynamics-implementation error [including the differentiation error of vector $b(t)$ as a part]. These errors may result from truncating/roundoff errors in digital realization and/or high-order residual errors of circuit components in analog realization (see [2, 4] and references therein). For the perturbed Z-type model (3) with model-implementation errors $\Delta_{m_1}(t)$ and $\Delta_{m_2}(t)$ involved, we could have the following theoretical results on robustness.

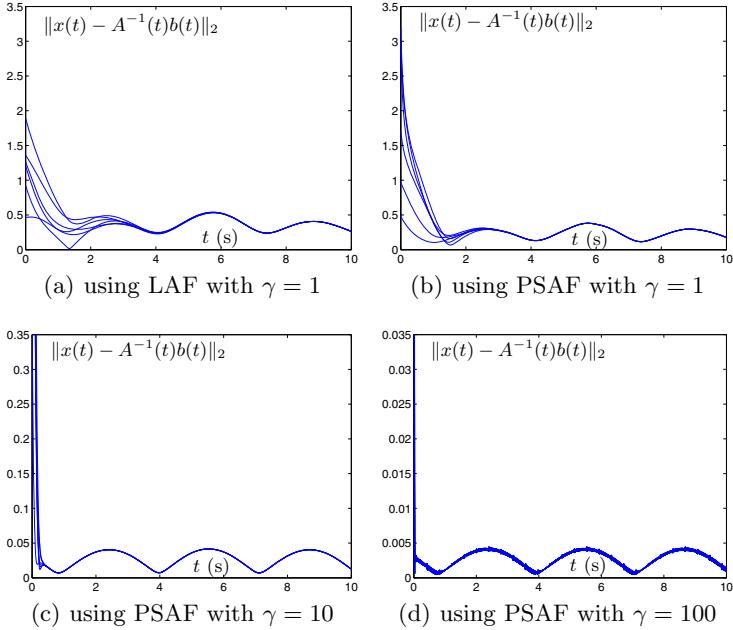


Fig. 1. Solution error $\|x(t) - A^{-1}(t)b(t)\|_2$ of perturbed Z-type ZLE model (3)

Theorem 1. If $\|\Delta_{m_1}(t)\|_F \leq \varepsilon_{m_1}$, $\|\Delta_{m_2}(t)\|_2 \leq \varepsilon_{m_2}$, $\|A^{-1}(t)\|_F \leq \varphi_A$ and $\|b(t)\|_2 \leq \varphi_b$ for any $t \in [0, \infty)$, and $0 < \varepsilon_{m_1}, \varepsilon_{m_2}, \varphi_A, \varphi_b < +\infty$, then the solution error $\|x(t) - A^{-1}(t)b(t)\|_2$ of the perturbed Z-type ZLE model (3) using the array of LAF or PSAF is upper bounded with the maximal steady-state error around $\sqrt{n}\varphi_A(\varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b)/(\gamma\rho - \varepsilon_{m_1}\varphi_A)$ under the design-parameter requirement $\gamma > \varepsilon_{m_1}\varphi_A/\rho$, where $\|\cdot\|_F$ and $\|\cdot\|_2$ denote respectively the Frobenius norm of a matrix and the two norm of a vector, and $\rho \geq 1$ exists. In addition, as $\gamma \rightarrow +\infty$, the steady-state computational error can be decreased to zero.

Proof. Let $\dot{e}(t)$ denote the time derivative of residual error $e(t) = e(x(t), t) = A(t)x(t) - b(t)$. Thus, the perturbed Z-type ZLE model (3) can be rewritten as

$$\dot{e} = -\gamma\mathcal{F}(e) - \Delta_{m_1}A^{-1}e + \Delta_{m_2} - \Delta_{m_1}A^{-1}b. \quad (4)$$

Then, a Lyapunov function candidate $v = \|e\|_2^2/2 = e^T e/2 = \sum_{i=1}^n e_i^2(t)/2 \geq 0$ (where T denotes the transpose of a matrix or a vector) for dynamic equation (4) could be defined. Evidently, v is positive definite ($v > 0$ for any $e \neq 0$ and $v = 0$ only for $e = 0$). Besides, if $\|e\|_2 \rightarrow \infty$, then $v \rightarrow \infty$. Furthermore,

$$\begin{aligned} \dot{v} &= e^T \dot{e} = e^T (-\gamma\mathcal{F}(e) - \Delta_{m_1}A^{-1}e + \Delta_{m_2} - \Delta_{m_1}A^{-1}b) \\ &= -\gamma e^T \mathcal{F}(e) + e^T Q e + e^T \Delta_{m_2} + e^T (-\Delta_{m_1}A^{-1}b) \\ &= -\gamma e^T \mathcal{F}(e) + e^T \frac{Q + Q^T}{2} e + e^T \Delta_{m_2} + e^T (-\Delta_{m_1}A^{-1}b) \end{aligned}$$

with $Q = -\Delta_{m_1} A^{-1}$. About the second term of the above equation, it follows from $\max_{1 \leq i \leq n} |\lambda_i(A)| \leq \|A\|_F$ that

$$\begin{aligned} e^T \frac{Q + Q^T}{2} e &\leq e^T e \max_{1 \leq i \leq n} \left| \lambda_i \left(\frac{Q + Q^T}{2} \right) \right| \\ &= e^T e \max_{1 \leq i \leq n} \left| \lambda_i \left(\frac{\Delta_{m_1} A^{-1} + (\Delta_{m_1} A^{-1})^T}{2} \right) \right| \\ &\leq e^T e \left\| \frac{\Delta_{m_1} A^{-1} + (\Delta_{m_1} A^{-1})^T}{2} \right\|_F \leq e^T e \|\Delta_{m_1}\|_F \|A^{-1}\|_F \leq e^T e \varepsilon_{m_1} \varphi_A. \end{aligned}$$

In addition, about the third and fourth terms of the above \dot{v} equation, it follows from $\max_{1 \leq i \leq n} |b_i| \leq \|b\|_2$ that

$$e^T \Delta_{m_2} \leq \sum_{i=1}^n |e_i| \max_{1 \leq i \leq n} |[\Delta_{m_2}]_i| \leq \sum_{i=1}^n |e_i| \|\Delta_{m_2}\|_2 \leq \sum_{i=1}^n |e_i| \varepsilon_{m_2},$$

and from the vector-matrix norms' relation $\|b\|_2 = \|b\|_F$ [6] that

$$\begin{aligned} e^T (-\Delta_{m_1} A^{-1} b) &\leq \sum_{i=1}^n |e_i| \max_{1 \leq i \leq n} |[\Delta_{m_1} A^{-1} b]_i| \leq \sum_{i=1}^n |e_i| \|\Delta_{m_1} A^{-1} b\|_2 \\ &\leq \sum_{i=1}^n |e_i| \|\Delta_{m_1} A^{-1}\|_F \|b\|_2 \leq \sum_{i=1}^n |e_i| \|\Delta_{m_1}\|_F \|A^{-1}\|_F \|b\|_2 \leq \sum_{i=1}^n |e_i| \varepsilon_{m_1} \varphi_A \varphi_b. \end{aligned}$$

Hence, in view of the above inequalities, we have

$$\begin{aligned} \dot{v} &\leq -\gamma e^T \mathcal{F}(e) + e^T e \varepsilon_{m_1} \varphi_A + \sum_{i=1}^n |e_i| \varepsilon_{m_2} + \sum_{i=1}^n |e_i| \varepsilon_{m_1} \varphi_A \varphi_b \\ &= -\sum_{i=1}^n |e_i| (\gamma f(|e_i|) - \varepsilon_{m_1} \varphi_A |e_i| - \varepsilon_{m_2} - \varepsilon_{m_1} \varphi_A \varphi_b). \end{aligned}$$

The following two situations could now be analyzed.

1. For time interval $[t_0, t_1]$, if $\gamma f(|e_i|) - \varepsilon_{m_1} \varphi_A |e_i| - \varepsilon_{m_2} - \varepsilon_{m_1} \varphi_A \varphi_b \geq 0, \forall i \in \{1, 2, \dots, n\}$, then $\dot{v} \leq 0$. By Lyapunov theory [3], the residual-error vector $e(t)$ of (4) converges towards zero [correspondingly, the state $x(t)$ of the perturbed Z-type model (3) converges towards the time-varying theoretical solution $x^*(t) = A^{-1}(t)b(t)$], as time t evolves until some time instant $t = \alpha$ such that $\dot{v} = 0$ [estimated on average $\gamma f(|e_i(\alpha)|) - \varepsilon_{m_1} \varphi_A |e_i(\alpha)| - \varepsilon_{m_2} - \varepsilon_{m_1} \varphi_A \varphi_b = 0, \forall i$, under requirement $\gamma f(|e_i(t)|) - \varepsilon_{m_1} \varphi_A |e_i(t)| \geq 0$].
2. For any time instant t , if $\gamma f(|e_i|) - \varepsilon_{m_1} \varphi_A |e_i| - \varepsilon_{m_2} - \varepsilon_{m_1} \varphi_A \varphi_b < 0, \exists i \in \{1, 2, \dots, n\}$, then perhaps $\dot{v} > 0$ and the residual-error vector $e(t)$ of (4) may thus not converge towards zero [correspondingly, the state $x(t)$ of the perturbed Z-type model (3) may not converge towards the theoretical solution $x^*(t) = A^{-1}(t)b(t)$ in this situation]. However, consider the worst case where $\dot{v} > 0$, $e(t)$ diverges outwards, and $e_i(t)$ increases: there exists time instant $t = \alpha$ such that $\dot{v} = 0$, estimated on average $\gamma f(|e_i(\alpha)|) - \varepsilon_{m_1} \varphi_A |e_i(\alpha)| - \varepsilon_{m_2} - \varepsilon_{m_1} \varphi_A \varphi_b = 0, \forall i$ [under requirement $\gamma f(|e_i(t)|) - \varepsilon_{m_1} \varphi_A |e_i(t)| \geq 0$].

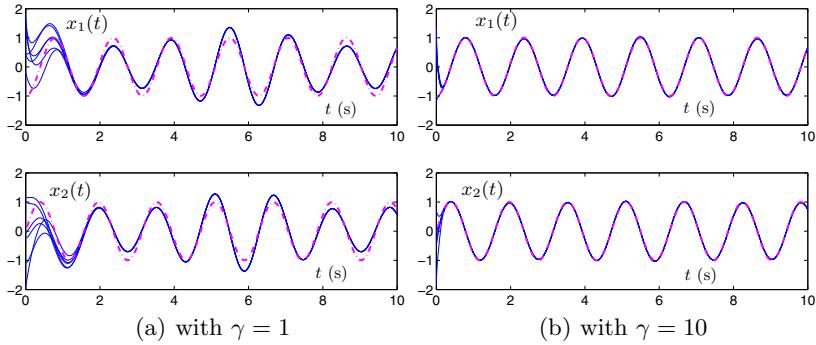


Fig. 2. γ -related robustness of perturbed Z-type model (3) using PSAF, where dashed-dotted curves correspond to theoretical Zhang quotient $A^{-1}(t)b(t)$ of ZLE (1), and solid curves correspond to the computed solutions with randomly-generated initial states.

In summary, in the worst case of $\dot{v} = 0$, the steady-state entry residual error is still upper bounded around $\bar{e}_i = (\varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b)/(\gamma\rho - \varepsilon_{m_1}\varphi_A)$ under design-parameter requirement $\gamma > \varepsilon_{m_1}\varphi_A/\rho$, $\exists\rho \geq 1$ satisfies $(\gamma\rho - \varepsilon_{m_1}\varphi_A)|e_i(\alpha)| = \varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b$ (see [3, 4] as well). Moreover, we have

$$\begin{aligned} \|x(t) - A^{-1}(t)b(t)\|_2 &= \|A^{-1}(t)(A(t)x(t) - b(t))\|_2 \\ &\leq \|A^{-1}(t)\|_F \|e(t)\|_2 \leq \varphi_A \sqrt{\sum_i^n e_i^2(t)} \leq \sqrt{n}\varphi_A \max_{1 \leq i \leq n} |e_i(t)|. \end{aligned}$$

Thus, it follows from average estimations that

$$\lim_{t \rightarrow \infty} \|x(t) - A^{-1}(t)b(t)\|_2 \lesssim \frac{\sqrt{n}\varphi_A(\varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b)}{(\gamma\rho - \varepsilon_{m_1}\varphi_A)}.$$

Evidently, the steady-state solution error of the perturbed Z-type ZLE model (3) can be made arbitrarily small by increasing the value of γ ; i.e., as $\gamma \rightarrow +\infty$, such a steady-state error of the perturbed Z-type model (3) decreases to zero. \square

Theorem 2. *In addition to robustness results in Theorem 1, perturbed Z-type ZLE model (3) possesses the following properties.*

1. *If LAF is used, then the steady-state entry residual error is upper bounded around $\bar{e}_i = (\varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b)/(\gamma - \varepsilon_{m_1}\varphi_A)$ under requirement $\gamma > \varepsilon_{m_1}\varphi_A$.*
2. *If PSAF is used, then we can remove the design-parameter requirement of γ being large enough, and superior robustness properties (e.g., faster convergence and smaller steady-state error) are achieved on the whole error range $e_i(t) \in (-\infty, +\infty)$, as compared to the LAF case.*

Proof. For the LAF case, the parameter $\rho \equiv 1$. From the proof of Theorem 1, we readily have $\bar{e}_i = (\varepsilon_{m_2} + \varepsilon_{m_1}\varphi_A\varphi_b)/(\gamma - \varepsilon_{m_1}\varphi_A)$ and the design-parameter requirement reduces to $\gamma > \varepsilon_{m_1}\varphi_A$.

For the PSAF case, the (steady-state) analysis has the following two parts.

1. For small error $|e_i(t)| \leq 1$, the sigmoid part of $f(\cdot)$ is activated with $f(|e_i(t)|) \geq |e_i(t)|$ and $\rho \geq 1$ (where the sign of equality is taken only for $|e_i(t)| = 1$). This, compared to $\rho \equiv 1$ in the linear-activation-function case, makes \dot{v} more negative (which implies a faster convergence) and generates \bar{e}_i smaller. In addition, the design-parameter requirement $\gamma > \varepsilon_{m_1} \varphi_A / \rho$ is relaxed with $\rho \geq 1$, as compared to the linear-activation case.
2. For large error $|e_i(t)| > 1$, the power part of $f(\cdot)$ is activated with $f(|e_i(t)|) = |e_i^p(t)| \geq |e_i(t)|$ and $\rho \geq 1$ (where the sign of equality is taken only for $|e_i(t)| = 1$). From the proof of Theorem 1, the original design-parameter requirement $\gamma f(|e_i(t)|) - \varepsilon_{m_1} \varphi_A |e_i(t)| \geq 0$ always exists, no matter how large ε_{m_1} and φ_A are. Thus, the parameter requirement on γ can be removed in the power-sigmoid situation. Besides, the amplifying effect of power activation function in this error range, $|e_i^p(t)| \geq |e_i(t)|$, makes \dot{v} more negative (which implies a faster convergence again) and generates \bar{e}_i smaller (due to $\rho = |e_i^{p-1}(\alpha)| \geq 1$ here), as compared to the LAF case.

Summarizing the above analysis, we know that, for the whole error range $e_i(t) \in (-\infty, +\infty)$ [except the isolated points $|e_i(t)| = 1, \forall i$, with equal performance], perturbed Z-type ZLE model (3) using the array of PSAF has superior robustness, as compared to the case of using the array of LAF. \square

4 Simulative Verification

Let us consider the following time-varying coefficients of ZLE (1):

$$A(t) = \begin{bmatrix} \sin(2t) & \cos(2t) \\ -\cos(2t) & \sin(2t) \end{bmatrix}, \quad b(t) = \begin{bmatrix} \sin(2t) \\ \cos(2t) \end{bmatrix},$$

and the perturbed Z-type model (3) with time-varying model-implementation errors as below (with $\varepsilon_{m_1} = \varepsilon_{m_2} = 0.5$):

$$\Delta_{m_1}(t) = \frac{\varepsilon_{m_1}}{\sqrt{2}} \begin{bmatrix} \cos(t) - \sin(t) \\ \sin(t) \cos(t) \end{bmatrix}, \quad \Delta_{m_2}(t) = \varepsilon_{m_2} \begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix}.$$

For comparison with the Z-type model's solution, the time-varying theoretical solution (i.e., the theoretical Zhang quotient) to ZLE (1) is given as

$$x^*(t) = A^{-1}(t)b(t) = \begin{bmatrix} -\cos(4t) \\ \sin(4t) \end{bmatrix}.$$

The computer-simulation results are shown in Figs. 1 and 2. As seen from Fig. 1, even with relatively large model-implementation errors, the solution error $\|x(t) - A^{-1}(t)b(t)\|_2$ of perturbed Z-type model (3) is still bounded and very small. In addition, as shown in Fig. 1(a) and (b), with $\gamma = 1$, the convergence time of perturbed Z-type model (3) using PSAF is faster than that using LAF. Furthermore, comparing Fig. 1(b)-(d), we see that, as design parameter γ increases from 1 to 10 and to 100, the convergence is evidently expedited and the steady-state solution error is decreased substantially (from around 0.45 to 0.045 and to 0.0045). Moreover, Fig. 2 shows the γ -related robustness of perturbed Z-type model (3) using PSAF, all substantiating well the theoretical results.

5 Conclusions

In this paper, by considering the model-implementation errors (i.e., the differentiation error and the dynamics-implementation error), the robustness properties of the perturbed Z-type model for ZLE solving have been investigated. As a result, even with relatively large model-implementation errors, the solution error of the perturbed Z-type model is still upper bounded. Besides, the convergence time and the steady-state residual error can be reduced effectively by increasing the value of design parameter γ . In addition, the perturbed Z-type ZLE model using PSAF has better robustness than that using LAF. Computer-simulation results have fitted well with the theoretical analysis, which have further demonstrated the excellent robustness of the Z-type model for ZLE solving.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (under Grants 61075121 and 60935001), and also by the Specialized Research Fund for the Doctoral Program of Institutions of Higher Education of China (with project number 20100171110045).

References

1. Steriti, R.J., Fiddy, M.A.: Regularized Image Reconstruction Using SVD and a Neural Network Method for Matrix Inversion. *IEEE Trans. Signal Process.* 41, 3074–3077 (1993)
2. Sturges Jr., R.H.: Analog Matrix Inversion. *IEEE Robot. Autom.* 4, 157–162 (1988)
3. Zhang, Y., Yi, C.: *Zhang Neural Networks and Neural-Dynamic Method*. Nova Science Publishers, New York (2011)
4. Zhang, Y., Ge, S.S.: Design and Analysis of a General Recurrent Neural Network Model for Time-Varying Matrix Inversion. *IEEE Trans. Neural Netw.* 16, 1477–1490 (2005)
5. Yi, C., Zhang, Y.: Analogue Recurrent Neural Network for Linear Algebraic Equation Solving. *Electr. Lett.* 44, 1078–1079 (2008)
6. Chen, Z., Sheng, J.: *An Introduction to Matrix Theory*. University of Aeronautics and Astronautics Press, Beijing (1998)

Appendix: Method Comparison

According to [3, 4], Z-type models and G-type models can be established for solving various time-varying problems of interest. As two parallel-processing dynamic systems, Z-type models and G-type models are different from each other. In this appendix, for making a clear comparison, as an example, design procedures of the Z-type model and G-type model for solving ZLE (1) are presented in Table 1. Furthermore, the main differences between Z-type models and G-type models for solving online time-varying problems are listed as follows.

Table 1. Z-type model design versus G-type model design for solving ZLE (1)

	Z-type	G-type
Error function	$e(x(t), t) = A(t)x(t) - b(t)$	$\mathcal{E}(x(t), t) = \ A(t)x(t) - b(t)\ _2^2/2$
Design formula	$de(x(t), t)/dt = -\gamma \mathcal{F}(e(x(t), t))$	$\dot{x}(t) = -\gamma \partial \mathcal{E}(x(t), t)/\partial x(t)$
Resultant model	$A(t)\dot{x}(t) = -\dot{A}(t)x(t) + \dot{b}(t)$ $-\gamma \mathcal{F}(A(t)x(t) - b(t))$	$\dot{x}(t) = -\gamma A^T(t)(A(t)x(t) - b(t))$

1. The design of Z-type models is based on the elimination of every entry of an indefinite matrix- or vector-valued error function (which could be positive, negative, bounded or even unbounded). In contrast, the design of G-type models is based on the elimination of a scalar-valued norm-type or square-type nonnegative energy function (which is at least bounded below).
2. Z-type models are depicted generally as implicit dynamic systems. In contrast, G-type models are depicted as explicit dynamic systems. Note that implicit dynamic systems may frequently arise in analog electronic circuits and systems. As compared to explicit dynamic systems, implicit dynamic systems have higher capabilities in representing dynamic systems.
3. Z-type models methodically and systematically exploit the time-derivative information of time-varying problems, and thus could be more effective on converging to the time-varying theoretical solutions of time-varying problems. In contrast, G-type models have not exploited such important information, and thus may be less effective on solving time-varying problems.
4. Z-type models could (globally) exponentially converge to the time-varying theoretical solutions of the problems of interest. By contrast, G-type models could only generate approximate passive results for theoretical solutions, lagging behind with much larger steady-state solution/residual errors.
5. Links exist between discrete-time Z-type models and Newton iterations (NI), while G-type models have not been reported to have an NI link.
6. Zhang fractals are yielded using the discrete-time complex-valued Z-type models to solve nonlinear equations in complex domain. Besides, Zhang fractals incorporate the well-known Newton fractals (generated by the Newton iteration) as special cases, though these two kinds of fractals are different. However, up to now, G-type or GD related fractal has not been reported, and is being researched by the authors of this paper.
7. Z-type models convert time-varying problems into linear or nonlinear equations to solve (i.e., zeroing). However, G-type models convert time-varying problems into minimization problems to solve (i.e., minimizing).
8. As for time-varying matrix inversion (or pseudoinversion), exact and complete links between Z-type models and the Getz-Marsden dynamic system (GMDS) have been discovered, with GMDS being a special case of Z-type models. However, G-type models do not have such GMDS links.
9. The derivation of Z-type models only requires the mathematical knowledge of B.S. level. By contrast, the derivation of G-type models may require much more complicated mathematical knowledge of M.S. or even Ph.D. level.

Orthogonal Waveform Design Based on the Modified Chaos Genetic Algorithm for MIMO Radar

Xin Fu, Xianzhong Chen, Qingwen Hou, Zhengpeng Wang, and Yixin Yin

School of Automation, University of Science and Technology Beijing, China
fxzn2006@163.com

Abstract. In view of the traditional genetic algorithm easily fall into local optimum in the late iterations, this paper puts forward an improved chaos genetic algorithm coded orthogonal signal design method which combines the chaos theory and genetic algorithm for MIMO radar. In order to prevent and overcome the ‘premature’ phenomenon in the process of optimization, the traversal features of the chaos optimization is introduced to the genetic algorithm, which reduces the autocorrelation peak side lobe and cross-correlation peak. Simulation results show that the proposed algorithm is feasible and effective.

Keywords: MIMO radar, orthogonal code signal, chaotic, genetic algorithm.

1 Introduction

MIMO radar transmitting orthogonal signal to replace the traditional radar transmitted coherent signal, each transmitting signal and the reception signal are independent and uncorrelated. MIMO radar systems require that every transmitting signal has small cross-correlation values and the autocorrelation function must have a narrow main lobe and lower side lobe feature to maximize the target structure information and space diversity gain and to achieve the high range resolution. So, waveform design is the principal important issue [1] for the MIMO radar system which will be successfully applied in target detection, high resolution imaging and other fields.

At present, there are many means for MIMO radar waveform design. For example, in paper [2], the simulated annealing algorithm is used in the design of the MIMO radar waveform by optimizing the cost function to get good signal correlation. In paper [3, 4], the traditional genetic algorithm is used to optimize orthogonal poly phase code to reduce the emission signal autocorrelation side lobe peak and cross-correlation peak. In paper [5] using good correlation performance of chaotic sequence, chaotic sequence is obtained by iteration method to improve the main lobe-to-side lobe ratio. Chaotic sequence was applied to design of random discrete frequency coded signal in paper [6] which can access to the need orthogonal signal. Compared with some optimization algorithms, there are still some gaps.

This paper will put forward the improved chaotic genetic algorithm for MIMO radar waveform optimization. The orthogonal poly phase code designed by using the method of this paper has lower autocorrelation side lobe peak and cross-correlation peak. This suggests that the method of this paper is feasible and effective.

1.1 MIMO Radar Transmitting Orthogonal Signal Model

Consider a MIMO radar system with L transmit antennas. Each unit transmits orthogonal coded pulse signal. Let $\{s_l(t)\}, l = 1, 2, 3, \dots, L$ denote the orthogonal coded pulse signal. Each signal has N of sub pulse duration for T_1 . Due to the orthogonality between signals, any two emission signal of cross-correlation function can be described by the following expression:

$$C(s_a, s_b, \tau) = \int_t s_a(t) s_b^*(t - \tau) dt = 0, a \neq b, \forall \tau \in R, a, b = 1, 2, 3, \dots, L \quad (1)$$

where $(\cdot)^*$ denotes the conjugate. To make signals with high resolution in range, signal's autocorrelation function can be described by the following expression:

$$A(s_l, \tau) = \frac{1}{E} \int_t s_l(t) s_l^*(t - \tau) dt = \begin{cases} 1, \tau = 0 \\ 0, others \end{cases}, l = 1, 2, 3, \dots, L \quad (2)$$

where E is the energy of signal $s_l(t)$.

The orthogonal poly phase coded pulse signal with M phase state has the following form:

$$\{s_l(n) = e^{j\varphi_l(n)}\}, n = 1, 2, 3, \dots, N, l = 1, 2, 3, \dots, L \quad (3)$$

where $\varphi_l(n) \in \left\{0, \frac{2\pi}{M}, 2 \cdot \frac{2\pi}{M}, \dots, (M-1) \cdot \frac{2\pi}{M}\right\}$ is the loading phase constant of the first n for signal l . Therefore, let S denotes the poly phase codes set,

$$S(L, N, M) = \begin{bmatrix} \varphi_1(1) & \varphi_1(2) & \cdots & \varphi_1(N) \\ \varphi_2(1) & \varphi_2(2) & \cdots & \varphi_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_L(1) & \varphi_L(2) & \cdots & \varphi_L(N) \end{bmatrix} \quad (4)$$

where N denotes the code length of the signal l and S is a matrix with $L \times N$.

The cross-correlation function and the autocorrelation function of phase coding sequence is given by:

$$C(s_a, s_b, k) = \begin{cases} \frac{1}{N} \sum_{n=1}^{L-k} s_a(n) s_b^*(n+k) = 0, 0 \leq k < N \\ \frac{1}{N} \sum_{n=-k+1}^N s_a(n) s_b^*(n+k) = 0, -N \leq k < 0 \end{cases} \quad a \neq b \quad (5)$$

$$A(s_l, k) = \begin{cases} \frac{1}{N} \sum_{n=1}^{L-k} s_l(n)s_l^*(n+k) = 0, & 0 < k < N \\ 1, & k = 0 \\ \frac{1}{N} \sum_{n=-k+1}^N s_l(n)s_l^*(n+k) = 0, & -N < k < 0 \end{cases} \quad (6)$$

where k denotes discrete time and $a, b, l = 1, 2, \dots, L$. If $|k| > N$, $C(s_a, s_b, k) = 0$,

$A(s_l, k) = 0$. The poly phase coded signals of MIMO radar is designed by the problem of combining optimization. Then we find out the requirements of the phase coding sequence as much as possible in accordance with equation (5) and (6). No matter adopting what kind of optimization algorithm, we design a reasonable objective function is an important prerequisite. The performance of the orthogonal signal can be evaluated by the autocorrelation function of side lobe peak and the peak level of the cross-correlation function. The objective function under the constraint of ASP and CP can be described as

$$\begin{aligned} E = & \lambda_1 \sum_{l=1}^L \max_{k \neq 0} |A(s_l, k)|^2 + \lambda_2 \sum_{a=1}^{L-1} \sum_{b=a+1}^L \max_{k \neq 0} |C(s_a, s_b, k)|^2 + \\ & \lambda_3 \sum_{l=1}^L \sum_{k=1}^{L-1} |A(s_l, k)|^2 + \lambda_4 \sum_{a=1}^{L-1} \sum_{b=a+1}^L \sum_{k=-(N-1)}^{N-1} |C(s_a, s_b, k)|^2 \end{aligned} \quad (7)$$

where $\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]$ is the weighted coefficient of the objective function. Then the orthogonal waveform design problem can be formulated as the following optimization problem: minimize (7) subject to (5) and (6) yields the quadrature phase sequence.

2 Waveform Optimization Algorithm

2.1 Improvement of the Tent Map

At present, generating chaotic sequence as the initial group by Logistic model or adding to the chaos random disturbance in mutation is adopted in most of the genetic algorithm. Chaotic variables are generated by the Logistic map

$$x_{k+1} = \infty * x_k * (1 - x_k), k = 0, 1, 2, \dots \quad (8)$$

where ∞ is the control parameter of chaotic state. When $\infty = 4$, the control system is in a state of complete chaos that is conducive to jump out of local optimum.

Tent map is known as the Tent map that is piecewise linear one-dimensional mappings. That is defined as follows:

$$x_{n+1} = \alpha - 1 - \alpha |x_n|, \alpha \in [1, 2] \quad (9)$$

when $\alpha=2$, that is referred to as the center of Tent map. It can be described as follows:

$$x_{k+1} = \begin{cases} 2x_k & 0 \leq x_k \leq 0.5 \\ 2(1-x_k) & 0.5 < x_k \leq 1 \end{cases} \quad (10)$$

We improve the Tent map by introducing method of stochastic equation [7]. The chaotic expression is defined as:

$$x_{k+1} = (1.99 * x_k \bmod 1) + \rho, \rho \in (0, 0.1) \quad (11)$$

Tent map can reach a small week point or fixed point and back into the chaotic state under the perturbation of stochastic equation. This process that is better be able to achieve global chaos optimization can enhance the ergodicity of the algorithm.

2.2 Orthogonal Poly Phase Coded Sequence Optimization Design

Genetic algorithm is a random search algorithm that is based on the principle of nature selection and genetics. With good global searching characteristics and better stability, it is especially fit for large and complex optimization. Chaos whose initial value has the properties of sensitivity, ergodicity and regularity can be used in the optimization problem. Genetic algorithm can not only search efficiently, but also can avoid falling into local optimum [8]. In recent years, many scholars put chaotic systems applying into the genetic algorithm. The optimization ability of genetic algorithm was improved greatly. We use Logistic model to generate chaotic sequence in the algorithm as the initial population, or add to the random disturbance of chaos in mutation to improve the performance of the algorithm. But there are still big blind search and slow convergence [9, 10]. This paper that is based on the improvement of Tent map adds to chaos disturbance for a generation of individuals in the group. That is to say, gene chaotic mutation operation can reduce the evolution algebra of genetic algorithm and is likely to produce the better gene sequences. It can not only improve the search speed of the algorithm but also avoids the problem of local convergence and premature effectively.

The genetic algorithm mainly simulates the evolution process of the biological search which is mainly done through crossover and mutation between chromosomes. We first need to binary encode orthogonal coding sequence group that simulated chromosome. Each phase state is mapped into a simple binary number. For example, the four phase coded signal $\varphi_l(n) \in [0 \ \pi/2 \ \pi \ 3\pi/2]$ is mapped into the simple binary number [00 01 10 11] which is one-to-one mapping. When

phase state M can't be divided exactly by 2, there will be $2^\delta - M$ redundant binary code. We select randomly an effective binary coded string to replace the redundant binary code. $\vec{X}(0)$ denotes the population of initial, $\vec{X}(i)$ denotes the population of the i th. The waveform optimization process with improved chaotic genetic algorithm is described as following (Fig. 1).

Step 1 Generate initial population and encode. The initial population $\vec{X}(0)$ is generated randomly with $i = 0$. Each set of the orthogonal poly phase coded sequence is as an individual. P denotes the size of each population. p_c denotes the crossover probability and p_m denotes the mutation probability. And the weighting value of w is set as: $\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [1, 1, 1, 1]$.

Step 2 Calculate the fitness value. Count the reciprocal of the objective function (7) as the fitness value. Calculate the fitness function value of every individual in the population as follows: $F(x) = 1/E(x)$, $F(x) \geq 0$. The higher the fitness value, the more easily inherit to the next generation.

Step 3 Judgment. Determine whether it is in line with the termination condition. If they meet the termination condition, computer could end of the algorithm. Then, the system could find out the best individual from the population. Then jump to Step4.

Step 4 Selection. According to the fitness value of each individual, we adopt the roulette method to select some excellent individuals from the i th generation of population $\vec{X}(i)$ that inherit to the next generation $\vec{X}(i+1)$, and generate new individuals to instead of those which are not be chosen.

Step 5 Crossing. The individuals in the population $\vec{X}(i)$ match in pairs randomly, and they replace or restructure parent individuals with crossover probability p_c to generate new individuals.

Step 6 Mutation. Each individual in population $\vec{X}(i)$ changes one or some genes with mutation probability p_m and selected efficient coding randomly to instead of redundant coding.

Step 7 Individuals whose fitness value is in the top 10% in mutated population $\vec{X}(i)$ are not do chaotic disturbance. But they will participate in the next genetic manipulation. Each individual parameter in the population is mapped to the Chaos Space $[0, 1]$ by rule: $y_{li} = x_{li}/(2\pi)$. where x_{li} is the individual parameter of the population $\vec{X}(i)$. Using the iterative formula of (11), we can calculate the chaotic vector after k iterations $y_{li}^{(k)}$, where k is used to identify the number of iterations in the chaotic sequence.

Step 8 Filter vectors after chaotic disturbance and calculate the new fitness value F'_{li} . If $F'_{li} \geq F_{li}$, let perturbed gene sequence instead of the original gene sequence, or to retain the original. The completion of the chaotic mutation process will generate mutated gene sequences.

Step 9 Judge the convergence of the fitness value of population $\vec{X}(i)$. If the sequence is convergent or the iteration number is equal to the designed maximum iteration number, the algorithm will be terminated, and the best solution so far will be what we need; else, jump to Step4.

3 Experimental Results and Analysis

We use the algorithm in this paper to design the transmitted waveform sets for MIMO radar. The main parameters are defined as follows: the population size $P = 100$; the crossover probability $p_c = 0.9$; the mutation probability $p_m = 0.1$; the biggest iteration step is 1000. Minimize (7) subject to (5) and (6) yields the quadrature phase sequence.

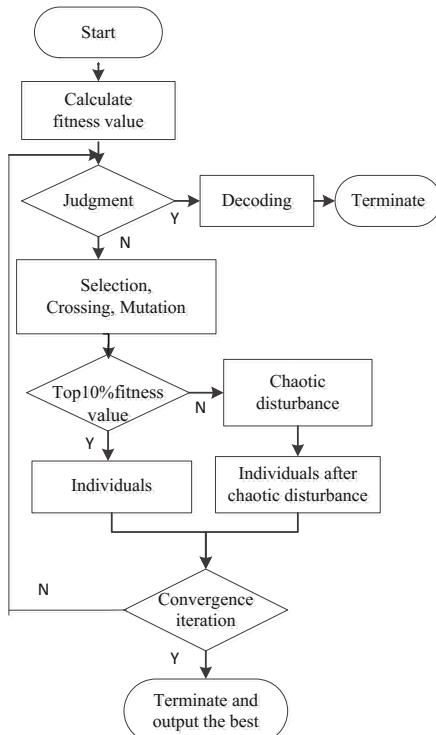


Fig. 1. Flow chart for CGA processing

Table 1. Optimization of the phase sequences

Code number	The phase sequence
Sequence1	0 1 0 2 2 1 1 1 3 3 2 3 3 3 3 1 0 2 3 3 2 1 3 3 2 2 0 1 0 0 0 2 3 1 1 3 0 3 2 0
Sequence2	2 2 3 1 1 1 1 1 2 1 0 2 2 2 2 1 3 2 0 2 3 3 1 1 2 1 3 3 1 0 1 3 2 0 3 3 0 3 0
Sequence3	1 0 3 2 1 2 0 2 0 2 3 3 2 2 0 1 0 3 1 0 2 1 2 0 3 1 0 2 1 0 2 2 2 2 1 0 3 2 3 0
Sequence4	3 1 1 3 1 2 0 3 3 0 0 2 0 0 2 1 3 3 1 3 1 3 2 3 0 0 2 2 3 2 1 0 3 2 0 2 0 3 2 0

Table 2. ASP and CP of the poly phase sequences

	Sequence1	Sequence2	Sequence3	Sequence4
Sequence1	0.1013	0.1773	0.1602	0.1332
Sequence2	0.1773	0.11	0.1521	0.1435
Sequence3	0.1602	0.1521	0.0886	0.1374
Sequence4	0.1332	0.1435	0.1374	0.0905

Table 3. Compare the results

Algorithm	The average ASP	The average CP
Standard genetic algorithm	-17.2dB	-13dB
Chaotic sequence	-19.2dB	-15.3dB
Algorithm in this paper	-20.21dB	-16.44dB

Table 1 shows a set of orthogonal poly phase coded sequence where $L=4$, $N=40$, $M=4$. The number multiplied by $\pi/2$ is the corresponding phase value. Table 2 shows autocorrelation side lobe peak (ASP) and cross-correlation peak (CP) of poly phase coded signals. The main diagonal is normalization of ASP, and the rest is normalization of CP. It can be conclude that the average ASP is about 0.0976(-20.21dB) and the average CP is about 0.1506(-16.44dB). Table 3 shows the average ASP and the average CP of the proposed algorithm have been improved significantly compared to the other two algorithms.

4 Conclusions

An effective algorithm with requirements imposed on both autocorrelation and cross-correlation functions has been developed for the design of orthogonal poly phase code sets used in MIMO radar for significantly improving radar performance. With the proposed optimization algorithm, some of the results are presented which have better correlation properties than available results designed by other algorithms.

Acknowledgments. This work is partially supported by National Science Foundation of Beijing under grant number 4132065.

References

1. Yang, Y., Blum, R.S.: MIMO Radar Waveform Design Based on Mutual Information and Minimum Mean-Square Error Estimation. *IEEE Transaction on Aerospace and Electronic Systems* 43, 330–340 (2007)
2. Deng, H.: Polyphase Code Design for Orthogonal Netted Radar Systems. *IEEE Trans. on Signal Processing* 52, 3126–3135 (2004)
3. Bo, L., Zishu, H., Jiankui, Z.: Polyphase Orthogonal Code Design for MIMO Radar Systems. In: Proc. of the International Conference on Radar, pp. 1–4. IEEE Press (2006)
4. Yu, Z., Jintao, S.: Phase-Coding Waveform Design for MIMO Radar Systems. *Acta Armamentarii* 1, 109–112 (2010)
5. Dong, S., Linrang, Z., Xin, L.: Polyphase Orthogonal Code Waveform Design Based on Chaotic Sequences for MIMO Radar. *Journal of Lanzhou University* 1, 100–106 (2011)
6. Yang, J., Qiu, Z.K., Xin, L.: Random Discrete Frequency Coding Signal Based on Chaotic Series. *Journal of Electronic & Information Technology* 33, 2702–2708 (2011)
7. Liang, S., Hao, Q., Jun, L.: Chaotic Optimization Algorithm Based on Tent Map. *Control and Decision* 20, 179–182 (2005)
8. El-Gohary, A., Al-Ruzaiza, A.S.: Chaos and Adaptive Control in Two Prey, One Predator System with Nonlinear Feedback. *Chaos, Solitons and Fractals* 34(2), 443–453 (2007)
9. Dongping, T.: Particle Swarm Optimization Algorithm Based on Tent Chaotic Sequence. *Computer Engineering* 36, 180–182 (2010)
10. Xiaohui, Y., Cheng, W.: A Novel Self-Adaptive Chaotic Genetic Algorithm. *Acta Electronica Sinica* 34, 708–712 (2011)

Automatic Object Tracking in Aerial Videos via Spatial-temporal Feature Clustering

Xiaomin Tong, Yanning Zhang, Tao Yang, and Wenguang Ma

School of Computer Science, ShaanXi Provincial Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University, Xi'an, China
xmtongnwpu@gmail.com, ynzhang@nwpu.edu.cn, yangtaonwpu@163.com

Abstract. Automatic detecting and tracking the objects from UAV videos is very important and challenging for both tactical and security applications. We present a robust object tracking system that is able to track multiple objects robustly in UAV videos. The main characteristics of the proposed system include: (1)A novel feature clustering based multiple objects tracking framework is proposed, which performs much better than the traditional foreground-blob-tracking-based methods. (2)Optical flow features are clustered both in spatial and temporal dimension to track multiple objects robustly even in the case of multiple objects cross moving. Extensive experimental results with quantitative and qualitative analysis demonstrate the robustness and effectiveness of our algorithm.

Keywords: Multiple objects tracking, optical flow, spatial-temporal trajectory clustering.

1 Introduction

With the increasing usage of UAVs for surveillance and other applications, it is of great interest to develop a fully automatic, efficient and robust object tracking system for UAV videos [1–5]. It can be widely applied in large area surveillance, search, rescue and traffic monitoring, especially for the non-cooperative targets tracking. For example, we can use it to track a car in the urban road, rescue a man in the wood or monitor some key places etc.

Many object tracking methods for UAV videos have been developed due to its various applications. Earlier typical attempt is the COCOA system [6]. It mainly contains three steps: stabilization [7–9], frame differencing, and blob tracking. However, it usually fails when the scene zooms due to the usage of Harris corner detection. Another prominent framework is proposed in [10] and [11], which performs iterative affine model estimation for image alignment, normal flow field for motion detection, graphs for representation and maintenance of a dynamic template of the moving objects. Although, this framework could achieve fast processing speed, it cannot handle the complex zooming scene. Aryo Ibrahim etc [12] construct the MODAT framework, using the SIFT feature [13] instead of Harris corner for frame matching. Unfortunately, all the above systems are under

the tracking-by-detection framework and usually fail in the complex surveillance scene. The challenges mainly come from the abrupt discontinuities in motion caused by the UAVs fast moving, low resolution noisy imagery, cluttered background, occlusion, significant change of scale, low contrast and small size of the target. All these make the detection result unstable so as to influence the tracking result. Besides, when multiple objects move crossing each other, tracking often fails due to the limitation of data association only in spacious dimension.

Recently, many object tracking algorithm basing on on-line learning theory [14, 15] are proposed. However, most of the state-of-the-art object tracking algorithms for UAV videos are usually started manually.

In this paper, we present a novel object tracking algorithm and system for real-time UAV videos. The system detects the motion trajectory by a KLT (Kanade-Lucas-Tomasi) tracker [16–18] rather than a background model. Thus, the algorithm is more robust to scale change, illumination change and other challenging cases thanks to the stability of optical flow feature. Then we confirm the object location by clustering the feature trajectories in both spatial and temporal dimension so as to track crossing objects robustly. Our system is fully automatic and does not require choosing object manually.

The rest of this paper is organized as follows. Section 2 introduces the framework of our system. Section 3 describes the details of our algorithm. Section 4 shows the experimental results and gives the discussions and Section 5 concludes the whole paper.

2 System Overview

The framework of the proposed system is shown in Fig. 1. There are mainly two modules: (1) a KLT-based feature tracking module for creation of the candidate feature point, tracking the candidate and existing feature point, and maintain or remove the feature point from the tracker list; and (2) a tracker clustering module for filtering the valid trackers, spatial trajectories clustering and temporal trajectories clustering. The individual components of the two modules are described in the following section.

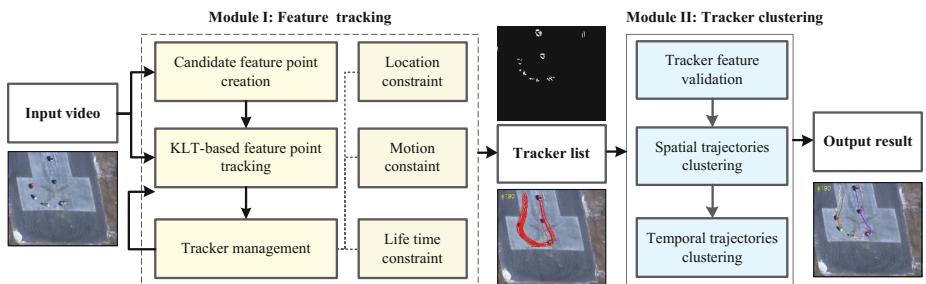


Fig. 1. Overview of our object tracking system

3 Spatial-temporal Clustering Based Tracking

3.1 KLT-Based Feature Tracking and Tracker Management

a) KLT based Optical flow

In order to track the object robustly in the challenging UAV videos, we adopt a KLT-based feature point tracking and clustering method instead of the background subtraction and blob tracking. For simplicity and robustness, we use the pyramidal implementation of the classical KLT tracker to get global motion vector. The core idea is to determine the local motion of window W from image I to image J and motion vector is calculated in the lowest level and propagated to higher resolution level by level. The number of pyramid levels is 5 and the patch size used is 5×5 pixels in our experiment for the trade-off between accuracy and efficiency.

b) Candidate feature creation

In an input frame, the optical flow of the good features located in the background scene accord with a transform model caused by the camera motion, while those belonging to moving object have distinguishing optical flow. Thus, we use RANSAC to get the global motion parameters and sort the candidate features that don't accord with the global motion. In our system, an affine model M is adopted to describe the global motion between two frames.

$$M = [R|T] \quad (1)$$

where R, T denote the rotation and translation parameter. Let $(x_t^i, y_t^i), (x_{t+1}^i, y_{t+1}^i)$ represent the location of feature i at time t and $t + 1$. Then, the projection error can be calculated as follow:

$$\text{Error} = \left\| (x_{t+1}^i, y_{t+1}^i)^T - M \cdot (x_t^i, y_t^i, 1)^T \right\|_2 \quad (2)$$

If Error is bigger than a given threshold, it is confirmed as a candidate feature.

By applying the pyramid KLT-based feature tracking and candidate feature point validation, an extensive feature point set with trajectories is obtained frame by frame. Meantime, we need to remove the redundant new trackers too close to the existing tracker and the wrong trackers with large motion vector.

3.2 Spatial-temporal Trajectories Clustering

As we have obtained lots of trajectories belonging to different objects, the main task is to classify them into multiple clusters. When objects are close to each other, it is difficult to obtain the objects trajectories merely via clustering the trajectories in spatial dimension. In a long term, these trajectories belonging to different object cannot be clustered into one cluster most times when objects are far away from each other, while those belonging to the same object should often have the same cluster label.

Proposition. Two trajectories belong to the same object with high probability only if they can be clustered into the same cluster consistently in consecutive frames.

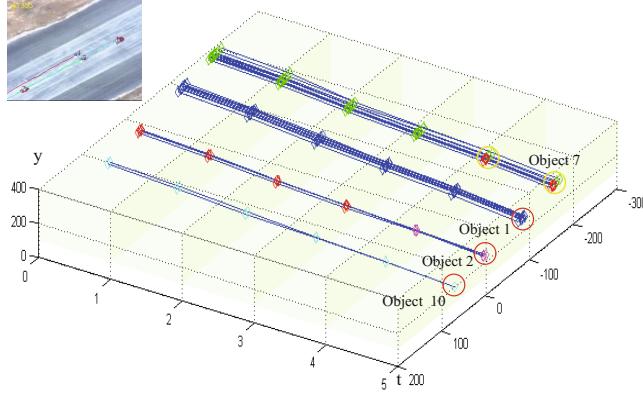


Fig. 2. Error in spatial feature clustering

For instance, in Fig. 2, there are lots of feature trajectories belonging to four objects in the red and yellow circle and the points with the same color at the same time means to belong to the same cluster. Fig. 2 shows that features belonging to object 7 are clustered to two different clusters in red and green only at time $t = 4$ and $t = 5$, while clustered into the same cluster in green at time $t = 0, \dots, 3$. Thus, these features belong to the same cluster according to the Proposition. Basing on the Proposition, we first cluster the trajectories in continuous N frames separately. Valid features are filtered if its length is not less than N . Let $F(i) = (f_1(i), f_2(i), \dots, f_t(i))$, $i = 1, 2, \dots, L$ denote all the L valid feature trajectories. In order to decide the cluster number in each frame, K-means clustering is used L times with clustering number $k = 1, 2, \dots, L$. With increasing of k close to the real object number, the within class variance decrease rapidly while decrease slowly when k increasing away from the real object number. Thus, we can find the turning point and obtain the cluster number K . Let $C(k, t)$ denote the center for cluster k and n_k feature points belong to cluster k .

$$C(k, t) = \frac{1}{n_k} \sum_{i=1}^{n_k} f_t(i) \quad (3)$$

where $\{f_t(i), i = 1, 2, \dots, n_k\}$ refers to the n_k feature points belonging to cluster k at time t . All these feature should satisfy the following constraint.

$$|f_t(i) - C(k, t)| < T_o \quad (4)$$

T_o refers to the distance threshold. For temporal clustering, we construct an association map AM with L columns and L lines. $AM(i, j)$ refers to the times $F(i)$ and $F(j)$ clustered in the same cluster in the continuous N frames. Then the maximal value in AM is N , meaning that the corresponding two features are clustered in the same cluster N times. Those small values usually refer to some wrong clustering when objects too close to each other. Thus we keep the correct association by equation (5).



Fig. 3. Our tracking result on EgTest01 dataset

$$A(i, j) = \begin{cases} 1 & \text{if } AM(i, j) > N * \alpha \\ 0 & \text{others} \end{cases} \quad (5)$$

where $\alpha \in [0, 1]$ denote the association factor. We can obtain the cluster result by $A(i, j)$ and track crossing objects robustly.

4 Experimental Results

To evaluate the performance of the proposed objects tracking algorithm, extensive practical tests were undertaken on public DARPA VIVID dataset (<http://vision.cse.psu.edu/data/vividEval/datasets/datasets.html>). The current C++ implementation of our algorithm runs on a 3.0GHz core 2 duo machine at the rate of 7 fps for 320*240 images without any optimization. Our tracker is a fully automatical system and we can't compare our algorithm with the state-of-the-art object tracking algorithms. This is because the current trackers such as Particle Filter [15], TLD [14] are usually started manually.

Fig. 3 shows our result on EgTest01 dataset, including 1821 frames with vehicles very similar to each other. Thus, blob-based tracker is prone to fail when vehicles pass others. However, our algorithm can deal with the crossing objects tracking due to our spatial-temporal clustering. In #1355 and #1515, vehicle 1 passes vehicle 2 and vehicle 10 separately and our tracker can track the vehicles robustly under this challenging condition.

Fig. 4 gives our tracking result on EgTest02 dataset with two sets of three civilian vehicles passing by each other on a runway. As we can see, vehicle 1, vehicle 2, and vehicle 3 are tracked continuously by our algorithm even with sudden change of scale between #700 and #910. Result on RedTeam dataset is shown in Fig. 5. The challenge of tracking this vehicle mainly comes from the long shadow(#498, #1800), change of scales(between #210 and #498, #1800



Fig. 4. Our tracking result on EgTest02 dataset



Fig. 5. Our tracking result on RedTeam dataset



Fig. 6. Our tracking result on PkTest01 dataset

and #1914) and so on. From Fig. 5 we can see that our algorithm could perform well in these complex scenes and track objects robustly. In Fig. 6, we give the result of PkTest01 dataset which is thermal IR data of a truck. In #1172, #1274 and #1337, the truck is passed by other vehicle separately and our algorithm could track the truck continuously even other vehicle is very close to the truck.

5 Conclusion

In this paper, we present a fully automatic object tracking system for aerial video. A KLT feature tracker is adopted to estimate the feature point trajectories and spatial-temporal feature clustering is applied to cluster the feature points into different objects. Objects are tracked automatically without starting manually. Extensive experimental results on large amount of test aerial videos illustrate the robustness and efficiency of our algorithm. In the future, we will concentrate on tracking the occluded object and dealing with other challenging cases in the UAV videos.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No.61231016, No.61272288), 2013 New People and New Directions Foundation of School of Computer Science in NPU (No.13GH014604), the NPU Foundation for Fundamental Research (No.JC201120, No.JC201148), and Plan of Soaring Star of Northwestern Polytechnical University (No.12GH0311).

References

1. Owen, M., Yu, H., McLain, T., Beard, R.: Moving ground target tracking in urban terrain using air/ground vehicles. In: GLOBECOM Workshops (GC Wkshps), 1816–1820 (2010)
2. Radhakrishnan, G.S., Saripalli, S.: Target tracking with communication constraints: An aerial perspective. In: IEEE International Workshop on Robotic and Sensors Environments (ROSE), pp. 1–6 (2010)
3. Zhu, S., Wang, D., Low, C.B.: Ground Target Tracking Using UAV with Input Constraints. *Journal of Intelligent & Robotic Systems* 1-4(69), 417–429 (2013)
4. Wang, J., Zhang, Y., Lu, J., Xu, W.: A Framework for Moving Target Detection, Recognition and Tracking in UAV Videos. *Affective Computing and Intelligent Interaction Advances in Intelligent and Soft Computing* 137, 69–76 (2012)
5. Fu, X., Feng, H., Gao, X.: UAV Mobile Ground Target Pursuit Algorithm. *Journal of Intelligent & Robotic Systems* 3-4(68), 359–371 (2012)
6. Ali, S., Shah, M.: Cocoa: Tracking in Aerial Imagery. *Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications III*, 62090D (2006)
7. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Alvey Vision Conference, vol. 15, p. 50 (1988)
8. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM* 24, 381–395 (1981)
9. Mann, S.: Compositing Multiple Pictures of The Same Scene. In: Proceedings of the 46th Annual IS&T Conference, vol. 2 (1993)

10. Cohen, I., Medioni, G.: Detecting and Tracking Moving Objects for Video Surveillance. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 319–325 (1999)
11. Medioni, G., Cohen, I., Bremond, F., Hong, S., Nevatia, R.: Event Detection and Analysis from Video Streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 873–889 (2001)
12. Ibrahim, A.W.N., Pang, C.W., Seet, G.L.G., Lau, W.S.M., Czajewski, W.: Moving Objects Detection and Tracking Framework for UAV-based Surveillance. In: Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp. 456–461 (2010)
13. Lowe, D.G.: Object Recognition from Local Scale-Invariant Features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1150–1157 (1999)
14. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In: Conference on Computer Vision and Pattern Recognition (2010)
15. Fan, Z., Li, M., Liu, Z.: An Improved Video Target Tracking Algorithm Based on Particle Filter and Mean-Shift. In: International Conference on Information Technology and Software Engineering, vol. 212, pp. 409–418 (2013)
16. Shi, J., Tomasi, C.: Good Features to Track. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
17. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. Carnegie Mellon University Technical Report (1991)
18. Yang, T., Zhang, Y., Shao, D., Li, Y.: Clustering Method for Counting Passenger Getting in a Bus with Single Camera. *Optical Engineering* 49(037203) (2010)

A Neural Network for Parameter Estimation of the Exponentially Damped Sinusoids

Xiuchun Xiao^{1,2}, Jian-Huang Lai², and Chang-Dong Wang³

¹ College of Information, Guangdong Ocean University, China

springxxc@163.com

² School of Information Science and Technology, Sun Yat-sen University, China

stsljh@mail.sysu.edu.cn

³ School of Mobile Information Engineering, Sun Yat-sen University, China

changdongwang@hotmail.com

Abstract. The problem of estimating the parameters of exponentially damped sinusoids (EDSs) has received very much attention in many fields. Strictly following the mathematic formulation of EDSs, we construct a specific neural network, termed EDSNN. In order to train EDSNN, a modified Levenberg-Marquardt iterative algorithm is derived. Profiting from good performance in fault tolerance of neural network, the proposed algorithm can be expected to possess a good performance in resistance to noise to some extent. Computer simulations have been conducted to apply this method to some EDSs signal models. The results substantiate the proposed EDSNN can obtain a higher precision for the parameters of the EDS component than the state-of-the-art algorithm.

Keywords: Exponentially damped sinusoids (EDSs), neural network, Levenberg-Marquardt algorithm, parameter estimation.

1 Introduction

Estimating the parameters of exponentially damped sinusoids(EDSs) is a very important task in many practical applications such as power system transient detection [1,2], speech analysis [3], etc. A number of techniques have been proposed to tackle this problem in the past. These techniques can be mainly classified as nonparametric [1,4–6], such as FFT [4,5] and wavelet analysis [6], and parametric ones [2,7], such as ESPRIT [2] and Prony analysis [8]. The nonparametric techniques are commonly computationally efficient and have less sensitivity to algorithm specified parameters. Unfortunately, these techniques often have inherent limitations such as suffering from frequency resolution or leakage effects in a unsynchronized sampling [4,9]. Compared with nonparametric methods, most parametric methods are model-based methods and can commonly achieve a relatively high accuracy [7].

In these years, parameter estimation using neural network methods is becoming popular for their high accuracy and good performance in resistance to noise [9]. Nevertheless, for the complicated nonlinear formulation of EDSs, neural network can not be directly applied in EDSs signal analysis.

In this paper, strictly following the mathematic formulation of EDSs signal, a specific topology of a neural network termed EDSNN for parameter estimation of EDSs has been constructed. It has three layers and the most important layer is the hidden-layer, which is mainly composed of some neurons and operating units and its mathematical formulas is consistent with the EDSs signal. So we can estimate the parameters of EDSs by training the EDSNN. In order to solve the weights of the EDSNN, a modified Levenberg-Marquardt algorithm(LM algorithm) is carefully derived according to the pre-defined objective function. Then the parameters of all the EDSs can be calculated from the weights of the converged EDSNN.

2 Proposed Approach

In this section, we will present the mathematical formulation of the practical problem firstly, then a neural network termed EDSNN is proposed according to it. In order to solve the weights of the proposed EDSNN, an adaptive learning algorithm based on improved LM is derived. At last, the parameters of each EDS component can be directly calculated using the converged weights of EDSNN.

2.1 Problem Formulation

In general, an actual signal consisting of n distinct exponentially damped sinusoid modes can be represented by their respective unknown damping factors, angular frequencies, amplitudes and initial phases. Assuming m samples drawn from the signal $y(t)$ with uniformly sampling interval time Δt are recorded as: $y(t_j) := y(j\Delta t)$, $j = 0, 1, 2, \dots, m - 1$, then, $y(t_j)$ can be formulated as follows:

$$y(t_j) = \sum_{i=1}^n A_i e^{\sigma_i t_j} \sin(\omega_i t_j + \varphi_i), \quad i = 1, 2, \dots, n, \quad j = 0, 1, \dots, m - 1, \quad (1)$$

where $y(t_j)$ denotes the signal sampled at t_j ; i denotes the mode order; A_i denotes the amplitude of mode i ; σ_i denotes the damping factor of mode i ; ω_i denotes the angular frequency of mode i ; φ_i denotes the initial phase of mode i ; n denotes the number of EDS components; m denotes the size of sample set.

By using the well-known equation: $\sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta)$, Eq.1 can be rewritten as:

$$y(t_j) = \sum_{i=1}^n (A_i e^{\sigma_i t_j} \sin(\omega_i t_j) \cos(\varphi_i) + A_i e^{\sigma_i t_j} \cos(\omega_i t_j) \sin(\varphi_i)) \quad (2)$$

$$i = 1, 2, \dots, n, \quad j = 0, 1, \dots, m - 1.$$

Thus, Eq.2 can be rewritten in a compact form as:

$$y(t_j) = \sum_{i=1}^n (w_i e^{v_i t_j} \sin(v'_i t_j) + w'_i e^{v_i t_j} \cos(v'_i t_j)), \quad (3)$$

where, $v_i = \sigma_i$, $v'_i = \omega_i$, $w_i = A_i \cos(\varphi_i)$, $w'_i = A_i \sin(\varphi_i)$, i and j are given in Eq.2. Obviously, the parameters σ_i , ω_i , A_i and φ_i of the i -th EDS component can be directly calculated as follows:

$$\sigma_i = v_i, \omega_i = v'_i, A_i = \sqrt{w_i^2 + (w'_i)^2}, \varphi_i = \arctan(w'_i/w_i). \quad (4)$$

2.2 Neural Network Model for Estimating the Parameters of EDSs(EDSNN)

In order to estimate the parameters of each EDS component in Eq.1, a corresponding specific neural network model is proposed. We call this specific neural network EDSNN. Fig.1 illustrates the topology of EDSNN. It has three layers, i.e., input-, hidden- and output-layers. The hidden-layer, which is mainly composed of $3n$ neurons and some operating units, is the most important part of EDSNN. The operating units can perform addition and multiplication operations. The $3n$ neurons can be separated into three classes and each class employs one of the three kinds of activation functions, i.e., exponential function, sine function or cosine function. As to the input- and output-layers, they are simple units and only need to perform some simple operations. For the convenience of expression, we can denote the weights between input- and hidden-layers as weight vector $\hat{v} := [\hat{v}_1 \hat{v}'_1 \hat{v}_2 \hat{v}'_2 \dots \hat{v}_n \hat{v}'_n]$ and the weights between hidden- and output-layers as weight vector $\hat{w} := [\hat{w}_1 \hat{w}'_1 \hat{w}_2 \hat{w}'_2 \dots \hat{w}_n \hat{w}'_n]$, respectively. The output is the weighted sum of the product of pairs of activation functions, denoted as, $\hat{y}(t_j)$.

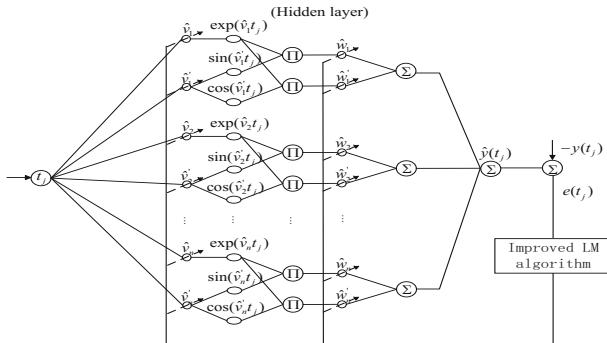


Fig. 1. Topology of EDSNN for estimating the parameters of EDSs signal

Therefore, the mathematical equation of EDSNN(Fig.1) can be expressed as:

$$\hat{y}(t_j) = \sum_{i=1}^n (\hat{w}_i e^{\hat{v}_i t_j} \sin(\hat{v}'_i t_j) + \hat{w}'_i e^{\hat{v}_i t_j} \cos(\hat{v}'_i t_j)), i = 1, 2, \dots, n. \quad (5)$$

Through observation and comparison, we can see that Eq.3 and Eq.5 are consistent in mathematical formulas. If $\hat{y}(t_j)$ in Eq.5 is regarded as the estimation value of $y(t_j)$ in Eq.3, then the weights \hat{w}_i , \hat{v}_i , \hat{w}'_i and \hat{v}'_i in Eq.5 can be regarded as an estimation of the parameters w_i , v'_i , w'_i and v'_i in Eq.3, respectively.

That is to say, the parameters of the i -th EDS component such as σ_i , ω_i , A_i and φ_i can be directly calculated from the weights \hat{w}_i , \hat{v}_i , \hat{v}'_i , \hat{w}'_i in EDSNN according to Eq.4. As a result, we can translate the estimation problem into a related optimization problem, i.e., how to construct a learning algorithm for training the EDSNN illustrated in Fig.1. In the next subsection, we will derive an improved LM algorithm to iteratively train the proposed EDSNN.

2.3 Improved Levenberg-Marquardt Algorithm

In order to estimate the parameters of each EDS component, we should update the weights in EDSNN to force it to approximate the EDSs signal. For this purpose, we can define a suitable objective function as follows:

$$E = \sum_{j=1}^m (\hat{y}(t_j) - y(t_j)). \quad (6)$$

where, $y(t_j)$ is the sampling value of the EDSs signal for $t = t_j$, $\hat{y}(t_j)$ is the actual output of EDSNN when its input is t_j .

Obviously, Eq. 6 is a non-linear optimization problem. In order to optimize the objective function E , we can define a function $F(\hat{\omega})$ as follows:

$$F(\hat{\omega}) = \hat{y}(t_j) - y(t_j) = \sum_{i=1}^n (\hat{w}_i e^{\hat{v}_i t_j} \sin(\hat{v}'_i t_j) + \hat{w}'_i e^{\hat{v}_i t_j} \cos(\hat{v}'_i t_j)) - y(t_j). \quad (7)$$

where, $\hat{\omega} := [\hat{v}_1 \hat{v}'_1 \hat{w}_1 \hat{w}'_1 \hat{v}_2 \hat{v}'_2 \hat{w}_2 \hat{w}'_2 \dots \hat{v}_n \hat{v}'_n \hat{w}_n \hat{w}'_n]$ is a vector containing all weights in the EDSNN. For all the inputs t_j , we have equations as follows:

$$\begin{cases} F_1(\hat{\omega}) = \sum_{i=1}^n (\hat{w}_i e^{\hat{v}_i t_1} \sin(\hat{v}'_i t_1) + \hat{w}'_i e^{\hat{v}_i t_1} \cos(\hat{v}'_i t_1)) - y(t_1) \\ F_2(\hat{\omega}) = \sum_{i=1}^n (\hat{w}_i e^{\hat{v}_i t_2} \sin(\hat{v}'_i t_2) + \hat{w}'_i e^{\hat{v}_i t_2} \cos(\hat{v}'_i t_2)) - y(t_2) \\ \vdots \\ F_m(\hat{\omega}) = \sum_{i=1}^n (\hat{w}_i e^{\hat{v}_i t_m} \sin(\hat{v}'_i t_m) + \hat{w}'_i e^{\hat{v}_i t_m} \cos(\hat{v}'_i t_m)) - y(t_m) \end{cases}. \quad (8)$$

In order to solve the weight vector $\hat{\omega}$, we can derive an improved LM algorithm as follows:

$$\hat{\omega}_{k+1} = \hat{\omega}_k + \Delta\hat{\omega}_k = \hat{\omega}_k - (J^T(\hat{\omega}_k)J(\hat{\omega}_k) + \mu_k I)^{-1} J^T(\hat{\omega}_k)F(\hat{\omega}_k), \quad (9)$$

where, $J(\hat{\omega}_k)$ is the Jacobi matrix of Eq.8, $\mu_k \in R$ is the learning rate defined as follows:

$$\mu_k := \alpha_k (\theta \|F(\hat{\omega}_k)\| + (1 - \theta) \|J^T(\hat{\omega}_k)F(\hat{\omega}_k)\|), \quad (10)$$

where, $\theta \in (0, 1)$, α_k is adjusted according to:

$$\alpha_{k+1} = \begin{cases} 4\alpha_k & \text{if } r_k < p_1 \\ \alpha_k & \text{if } r_k \in [p_1, p_2] \\ \max(\alpha_k/4, \tau) & \text{else} \end{cases}. \quad (11)$$

where, $0 < p_0 < p_1 < p_2 < 1$, $\tau > 0$, r_k is defined as follows:

$$r_k = \frac{Ared_k}{Pred_k}, \quad (12)$$

where, $Ared_k := \|F(\hat{\omega}_k)\|_2^2 - \|F(\hat{\omega}_k + \Delta\hat{\omega}_k)\|_2^2$, $Pred_k := \|F(\hat{\omega}_k)\|_2^2 - \|F(\hat{\omega}_k + J(\hat{\omega}_k)\Delta\hat{\omega}_k)\|_2^2$.

It is worth mentioning that the improved LM algorithm has better convergence compared to gradient descent method and traditional LM algorithm [9,10]. Thus, by training the EDSNN with the improved LM algorithm, we can achieve higher accuracy, which will be further substantiated in the section 3.

3 Simulation Verification

In this section, we perform numerical experiments to demonstrate the effectiveness of the proposed EDSNN approach. Two cases are considered to test the proposed method. It is worth further pointing out that, for both the simulation experiments in this section, the values of all the relevant parameters of EDSNN are fixed, which are listed as follows:

$$\tau = 10^{-8}, \theta = 0.5, \alpha_1 = 0.1 + \tau, p_0 = 10^{-4}, p_1 = p_0 + 0.25, p_2 = p_1 + 0.5.$$

3.1 Case Study I

The proposed EDSNN is firstly applied to estimate parameters of a simple signal which contains only one EDS component given in the literature [2]. The samples used for parameter estimation are generated using the following equation:

$$y(t) = 2.0e^{-2.5t} \sin(2\pi \times 5t + 3.0), \quad (13)$$

where, $A_1 = 1.0$, $\sigma_1 = -0.025$, $\omega_1 = 2\pi \times 0.4$, and $\varphi_1 = 3.0$. The noise-free and noisy versions of this artificial swing curve are shown in Fig.2(a).

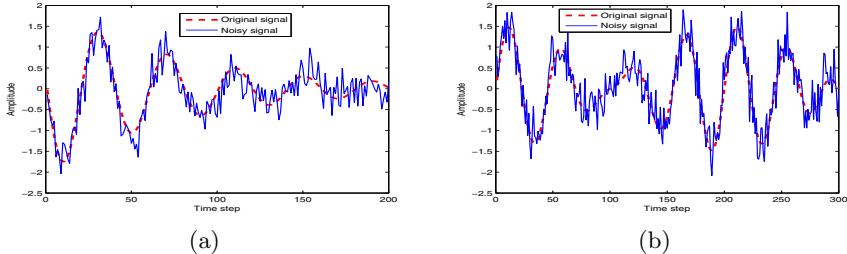


Fig. 2. Noise free and noisy versions: (a) a simple signal which contains only one EDS component [2], (b) a relative complex signal which contains two EDS components [1]

Noise Free Case. To obtain a high precision for the parameter estimation of the signal in Eq.13, we can set the objective error ε be very low, e.g., $\varepsilon = 10^{-20}$, and randomly generate the initial weights of EDSNN. In order to force EDSNN to converge, we update its weights by using the improved LM algorithm, then, the parameters of the signal denoted in Eq.13 can be estimated by using Eq.4. Table 1 illustrates the actual parameters of the original signal, the estimated results and the relative error by using the proposed EDSNN.

As shown in Table 1, we can see that the proposed EDSNN is capable of obtaining the parameters of the signal containing only one EDS component with almost no error.

Table 1. The actual, estimated parameters and relative error for signal denoted in Eq.13 with no noise

Parameters	A_1	σ_1	f_1	φ_1
Actual	2.0	-2.5	5.0	3.0
Estimated	2.0000	-2.5000	5.0000	3.0000
Relative Error	0.0000	0.0000	0.0000	0.0000

With Noise Case. In practical applications, noise disturbance is commonly an inevitable problem. In this part, we will consider the case when different signal-to-noise ratio (SNR) levels of white Gaussian noise (WGN) is added to the simulation signal denoted by Eq.13. In order to validate the robustness of the proposed EDSNN, we test and compare it with the popular ESPRIT method [2] in a variety of levels of SNRs. Table 2 illustrates the actual parameters of the original signal and the estimated results by using the proposed EDSNN.

Table 2. The actual, estimated parameters, and the relative error for signal denoted in Eq.13 with a variety of levels of SNRs

Actual	Methods	Estimated				Actual	Methods	Estimated			
		30dB	20dB	10dB	5dB			30dB	20dB	10dB	5dB
A_1 =2.0	ESPRIT	1.999	2.033	2.110	2.073	f_1 =5	ESPRIT	4.999	4.984	5.005	5.154
	EDSNN	2.003	2.001	1.991	2.004		EDSNN	4.998	5.000	5.000	5.000
σ_1 =-2.5	ESPRIT	-2.492	-2.277	-2.606	-2.812	φ_1 =3	ESPRIT	3.004	3.027	2.966	2.939
	EDSNN	-2.506	-2.496	-2.490	-2.518		EDSNN	3.002	2.999	2.999	2.998

As shown in Table 2, we can see that all of the estimation results by EDSNN have very high precision even the signal-to-noise ratio decaying to 5dB. Beyond that, we can also see that all of the estimation results by the EDSNN are better than the ESPRIT method [2] except for SNR=30dB. All these results show that our method is more robust to the noise ratio than ESPRIT [2].

3.2 Case Study II

In this subsection, the proposed EDSNN is applied to estimate parameters of a relatively complex signal which contains two EDS components given in the literature [1]. The samples used for parameter estimation are generated by using the following equation:

$$y(t) = 1.0e^{-0.025t} \sin(2\pi \times 0.4t) + 0.5e^{0.037t} \sin(2\pi \times 0.5t), \quad (14)$$

where, $A_1 = 1.0$, $\sigma_1 = -0.025$, $\omega_1 = 2\pi \times 0.4$, $A_2 = 0.5$, $\sigma_2 = 0.037$, $\omega_2 = 2\pi \times 0.5$, and the initial phase φ_1 and φ_2 are both assumed to equal to zero in this case. This artificial swing curve is shown in Fig.2(b).

Noise Free Case. Similar to case study I, we set the objective error $\varepsilon = 10^{-20}$, and randomly generate the initial weights of EDSNN. Then, we train EDSNN by using the improved LM algorithm and estimate the parameters by using Eq.4. Table 3 illustrates the actual parameters of the original signal, the estimated results and the relative error by using the proposed EDSNN.

Table 3. The actual, estimated parameters and relative error for signal denoted in Eq.14 with no noise

Parameters	Mode 1			Mode 2		
	A_1	σ_1	f_1	A_2	σ_2	f_2
Actual	1.0	-0.025	0.4	0.5	0.037	0.5
Estimated	1.0000	-0.0250	0.4000	0.5000	0.0370	0.5000
Relative Error	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

As shown in Table 3, we can see that the proposed EDSNN can obtain the parameters of the signal containing two EDS components with almost no error.

With Noise Case. In order to validate the robustness of the proposed EDSNN, we test and compare it with the popular ESPRIT method [2] in a variety of levels of SNRs. Suppose simulation signal II is polluted with different levels of Gaussian white noise. Table 4 illustrates the actual parameters of the original signal and the estimated results by using the proposed EDSNN.

Table 4. The actual, estimated parameters, and the relative error for signal denoted in Eq.13 with a variety of levels of SNRs

Actual	Estimated				Actual	Estimated			
	30dB	20dB	10dB	5dB		30dB	20dB	10dB	5dB
$A_1 = 1.0$	0.9998	1.0009	1.0113	1.0049	$A_2 = 0.5$	0.4993	0.5038	0.4987	0.4981
$\sigma_1 = -0.025$	-0.0249	-0.0249	-0.0257	-0.0264	$\sigma_2 = 0.037$	-0.3674	-0.0361	-0.0374	-0.0373
$f_1 = 0.4$	0.4000	0.4000	0.4000	0.3999	$f_2 = 0.5$	0.5000	0.5000	0.5000	0.5000
$\varphi_1 = 0$	0.001	-0.002	-0.001	-0.001	$\varphi_2 = 0$	-0.0248	-0.0249	-0.0256	-0.0263

As shown in Table 4, we can see that all the results by the proposed EDSNN have very high precision for estimating the relative complex signal contains two EDS components, even the signal-to-noise ratio decaying to 5dB. On the other hand, the estimating results of ESPRIT method are very unstable and also far away from the actual values, so we cannot list its results in Table 4. As a result, when a signal contains many EDS components in low level of SNR, we can consider the EDSNN instead of ESPRIT method for estimating its parameters.

4 Conclusions

Generally speaking, neural network cannot be directly applied to parameter estimation of EDSs. For the purpose of solving this problem, we construct a specific topology of a neural network termed EDSNN strictly following the mathematic formulation of EDSs signal. Then we derive an improved LM algorithm to train EDSNN. At last, the parameters of EDSs can be estimated according to the weights of EDSNN. Profiting from the strictly consistency of the mathematical formulas between the EDSNN and the EDSs model, and also profiting from the good performance of the improved LM algorithm, the proposed algorithm can achieve a very high precision and have good performance in resistance to noise.

Computer simulation results substantiate the proposed EDSNN can obtain a higher precision for damping factors, frequencies, also amplitudes and initial phases of all the EDS components than the state-of-the-art algorithm for noise free or noisy case. As a result, when a signal contains many EDS components in low level of SNR, or high precision is needed, we can consider the EDSNN instead of ESPRIT method for estimating its parameters.

Acknowledgements. This work was supported by NSFC (61173084 and 61128009), National Science & Technology Pillar Program (No:2012BAK16B06).

References

1. EL-Naggar, K.M.: On-line measurement of low-frequency oscillations in power systems. *Measurement* 42(5), 716–721 (2009)
2. Najy, W.K.A., Zeineldin, H.H., Kasem Alaboudy, A.H., Woon, W.L.: A bayesian passive islanding detection method for inverter-based distributed generation using esprit. *IEEE Transactions on Power Delivery* 26(4), 2687–2696 (2011)
3. Jensen, J., Heusdens, R., Jensen, S.H.: A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids. *IEEE Transactions on Speech and Audio Processing* 12(2), 121–132 (2004)
4. Agrež, D.: A frequency domain procedure for estimation of the exponentially damped sinusoids. In: I2MTC 2009, pp. 1321–1326 (2009)
5. Qian, H., Zhao, R., Chen, T.: Interharmonics analysis based on interpolating windowed fft algorithm. *IEEE Transactions on Power Delivery* 22(2), 1064–1069 (2007)
6. Barros, J., Diego, R.I.: Analysis of harmonics in power systems using the wavelet-packet transform. *IEEE Transactions on Instrumentation and Measurement* 57(1), 63–69 (2008)
7. Sun, W., So, H.: Accurate and computationally efficient tensor-based subspace approach for multi-dimensional harmonic retrieval. *IEEE Transactions on Signal Processing* 60(10), 5077–5088 (2012)
8. Feilat, E.A.: Prony analysis technique for estimation of the mean curve of lightning impulses. *IEEE Transactions on Power Delivery* 21(4), 2088–2090 (2006)
9. Xiao, X., Jiang, X., Xie, S., Lu, X., Zhang, Y.: A neural network model for power system inter-harmonics estimation. In: BIC-TA 2010 (2010)
10. Chen, P.: Why not use the levenberg–marquardt method for fundamental matrix estimation? *IET Computer Vision* 4(4), 286–294 (2010)

Local-Global Joint Decision Based Clustering for Airport Recognition

Bingxin Qu, Yanning Zhang, and Tao Yang

School of Computer Science

ShaanXi Provincial Key Laboratory of Speech and Image Information Processing
Northwestern Polytechnical University, Xi'an, China
qubingxin@mail.nwpu.edu.cn, ynzhang@nwpu.edu.cn, yangtaonwpu@163.com

Abstract. Airport recognition plays an important role in many computer vision tasks. This paper proposes a novel method for airport recognition including the generation of clustering threshold adaptively and automatically based on local-global joint decision. The novelties of the approach include: (1) a new adaptive clustering framework we called local-global joint decision based clustering is designed instead of a fixed threshold. (2) Within the special framework, we propose a statistical probability model based on a new local feature as the preliminary decision strategy which provide local information. (3) Another part of the framework we called reconfirmation strategy contains an innovative relative similarity model that uses global information to estimate the confidence coefficient of the result from preliminary decision. Extensive experimental results demonstrate the feasibility of our approach. In addition, our method outperforms the related algorithms in terms of recognition accuracy under a variety of situations.

Keywords: Airport recognition, local-global joint decision, adaptive clustering.

1 Introduction

With the gradual development of the unmanned aerial vehicle(UAV), a growing number of researchers focus on the automatic detection and recognition of typical target such as airport in aerial images. There exist a variety of effective recognition algorithms[1][2][3] for some specific background through the joint effort of scholars. Unfortunately, there has been no algorithm which is universal and expandable, see Fig.1(b)(c). Previous methods simplify this problem with the hypothesis that almost no straight lines will seriously interfere with the effect of recognition. In the case of wide application, the method under the assumption has weak noise immunity and make simultaneously the risk of false recognition incomparable.

This paper proposes a novel system that can make the airport recognition system have better noise immunity and a good ability of robustness. Inspired by the work of [4], an idea of adaptive clustering based on local-global joint decision is produced to reduce the influence of other straight lines. Rather than

specify a fixed clustering threshold, we scan a range of thresholds and calculate thresholds according to different situations automatically via a local-global joint decision based clustering. The novel clustering method includes a preliminary decision by a probability statistical model based on a new local feature[5] and a reconfirmation via a relative similarity model means global information. The result is an automatic algorithm that avoids troubles like the risk of the misrecognition because of wrong clustering. Additionally, it could significantly improve airport recognition's automation and intelligence in order to minimize human involvement.

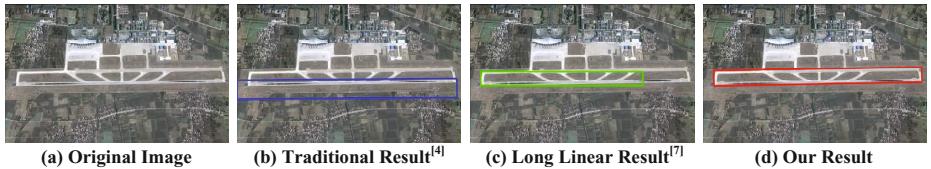


Fig. 1. Traditional airport recognition algorithm result, long linear result and our result

The main contributions of our paper are as follows: (1) we design a new adaptive clustering framework which is determined by the local-global joint decision that includes preliminary decision and reconfirmation, when traditional airport recognition methods could not change the clustering threshold depending on the sundry images. (2) The paper proposes a new preliminary decision strategy based on a probability statistical model which uses a novel local feature, which could provide special information for later process. (3) By observing some output result of preliminary decision strategy, we model a relative similarity model means global information of the reconfirmation strategy to estimate the confidence coefficient of the result from preliminary decision strategy. (4) We design an improved automatic airport recognition system which could handle clustering threshold via our neoteric joint decision part automatically for multifarious situations without human participation.

The rest of the paper is structured as follows: we will discuss a brief summary of the related work in this area in the next section. In section 3, we will present our adaptive clustering threshold framework via the local-global joint decision clustering and propose the two improved parts: a probability statistical model and a relative similarity model. Section 4 shows the experimental results include the comparison among those image segmentation methods and our novel joint decision-making system. Finally, the conclusions of our paper will show in section 5.

2 Related Work

In this section, we will summarize some work that relate to the research which we build upon in developing our algorithm, including airport recognition and adaptive clustering threshold.

Airport Recognition. In recent years, there exist various airport recognition algorithms for different situations. Most of the algorithms are based on the runway detection. Runway is the most typical marker which not only is inevitable existing in the airport, but also has some very obvious geometric characteristics. However, most of the algorithms described above are still not satisfactory because they addressed the problem of airport recognition have many limitations of the environment settings. In 2009, Hao[6] put forward a method which included improved Hough transform and morphological processing for airport recognition in complex environments. Soon after, Weng[4] skillfully took advantage of image segmentation to reduce the complex environment interference, and the straight lines of runway will be gained after Canny operator edge extraction and Hough transform, and then the Zernike moment and some geometric characteristics make up of an airport feature that will input into the SVM classifier for confirmation. In [1], an airport was described by a set of improved scale-invariant feature transform key-points. In 2012, Cao[7] proposed a runway recognition method based on long linear characteristics.

Adaptive Clustering Threshold. The K-means algorithm may be one of the most popular clustering methods. More and more scholars realized the fixed clustering threshold which need users provide in advance has no self-adaptability. Tapas Kanungo[8] analyzed and implemented an efficient K-means clustering algorithm in 2002. Li[9] proposed a novel threshold cutting method instead of long side cutting method[10], that improved traditional K-means algorithm. In this paper, we intend to build a new joint decision-making clustering which includes two models, a probability statistical model and a relative similarity model, to determine the threshold of distance of two lines automatically and intelligently.

3 Local-Global Joint Decision Based Clustering

Our airport recognition algorithm is similar with the methods we described above in spirit. Where our work differs are as follow: Rather than the fixed clustering threshold, we design a local-global joint decision to generate the clustering threshold adaptively. Specifically, our algorithm proceeds as the Fig.2 shows, which mainly includes three parts: runway detection, local-global joint decision based clustering and the SVM recognition. In the first module, by gaining the binary image through image segmentation after the image preprocessing, we extract the edge information of the binary image for straight lines detection. Then the Hough transform can be used to output the image contains straight lines. Runway extending and contact way searching proceed afterwards. In a word, the output of the first module is an image masked by runway. In the second module, with the image contains runways as our input, local-global joint decision based clustering is initiated. By lines clustering based on distance of two lines we can get a set of undetermined airport regions in accordance with different clustering thresholds. The best undetermined airport region with the highest confidence coefficient will be output when it both through the two parts of joint decision

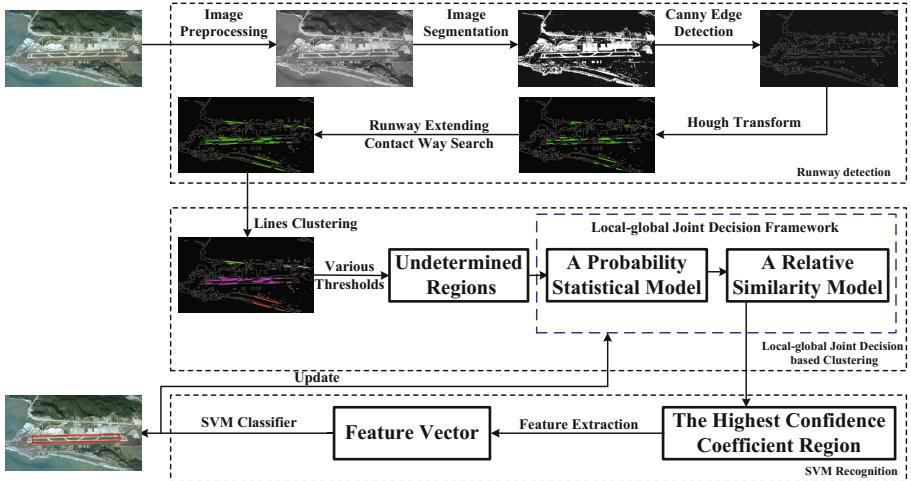


Fig. 2. Airport recognition algorithm of the local-global joint decision based clustering

clustering. The flashpoint must be the design of two novel models from the joint decision clustering; we will represent them in 3.1 and 3.2 in detail. Finally, the best undetermined airport region input into the third module which includes SVM classifier and feature extraction such as HOG, then output the final result contains an airport. The algorithm results are deferred to section 4.

3.1 Statistical Probability Model Based on a New Local Feature

As the first module of our framework, the algorithm of building the novel statistical probability model which provides local information is showed in Fig.3.

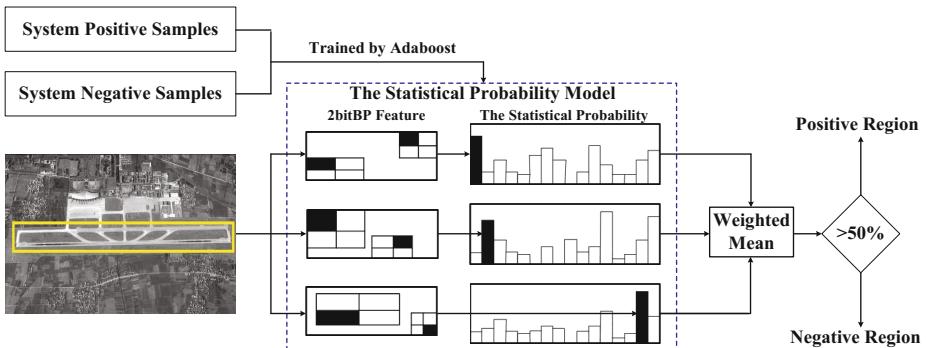


Fig. 3. The statistical probability based on 2bitBP feature

The statistical probability model is trained by Adaboost algorithm from system positive and negative samples, where system positive samples contain a lot of various airport regions and system negative samples contain other region that there is no airport. A new local feature means the 2bitBP feature[5]. And in particular the N 2bitBP features will not change unless the system positive and negative samples changed. The judging criterion of positive and negative is given as follows:

$$P_{positive} = \sum_{i=1}^N (flag \bullet W(i)), \begin{cases} flag = 1, if(SP(i) >= 0.5) \\ flag = -1, if(SP(i) < 0.5) \end{cases} \quad (1)$$

Where $P_{positive}$ is the possibility of the undetermined region that is treated as a positive one. SP denotes the whole statistical probabilities, $SP(i)$ is the i-th of the statistical probabilities. And then $W(i)$ is the weight of the i-th statistical probability which is produced in the process of Adaboost training. If $P_{positive}$ beyonds the half of the sum of all the weights, the region is positive; otherwise it is negative.

3.2 Relative Similarity Model

In relative similarity model, there are two main components including relative similarity positive samples and relative similarity negative samples. With those system positive and negative samples, the relative similarity model uses global image information could be trained using the following steps:

1. Arrange the sets of system positive samples and system negative samples randomly.
2. Process each sample in the ordering system sample sequence through the following actions: after computing the relative similarity that we called confidence coefficient between one sample and other samples in system sample sequence, there are two processes for us to choose; one situation is that if the predictive value of the one sample is true, the relative similarity is less than a threshold at the same time, this sample should be added to the set of relative similarity positive samples; on the flip side, the predictive value is false, but the relative similarity beyond a certain threshold, the sample must be added to the set of relative similarity negative samples.
3. Output both the relative similarity positive samples and relative similarity negative samples.

$$Conf_{RS} = \frac{Ndistance}{Ndistance + Pdistance} \quad (2)$$

$$Ndistance = 1 - \max(Nnccdistance) \quad (3)$$

$$Pdistance = 1 - \max(Pnccdistance) \quad (4)$$

In order to compute the relative similarity, a number of variables should be introduced at first. Define the relative similarity $Conf_{RS}$ as above. Where $Nnccdistance$ is the distance between one sample and a set of negative samples and $Pnccdistance$

is the distance between one sample and a set of positive samples. The distance here means a distance based on normalized cross correlation, we get:

$$Nnccdistance_k = \frac{1}{2} \bullet \left(\frac{corr}{\sqrt{norm_1 \bullet norm_2}} + 1 \right) \quad (5)$$

$Nnccdistance_k$ is distance between one sample and the k-th sample in negative samples, $corr$ means their covariance while $norm_1$ and $norm_2$ are their modules. All the samples including system positive and negative samples and the undetermined region must be normalized into a uniform size. The same processing is for the positive.

The entire whole relative similarity model has been generated. Given an undetermined region, compute the relative similarity based on the relative similarity positive and negative samples after normalization and then if the relative similarity is less than a fixed threshold, that region won't be masked; otherwise, it will be masked as a positive one.

4 Experimental Results

In order to evaluate the performance of the proposed approach, we have set up a series of experiments below. Experiments have been conducted using our own database from Google Earth, includes various structure, scale change, rotation and so forth. Details of the result as well as the analysis of the algorithm will be given in the following subsections.

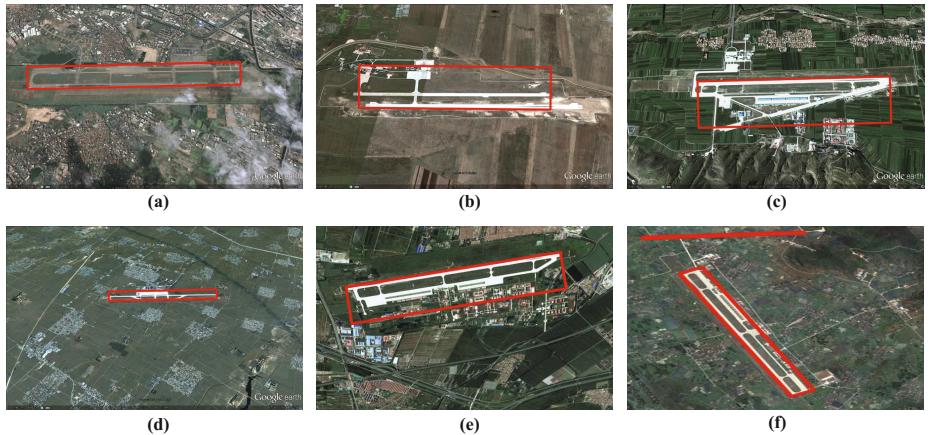


Fig. 4. Airport recognition results under different conditions

4.1 Results and Analysis of Our Method

We provide the airport recognition results of parts of the test images in Fig.4, where the airport region are overlain on the original images and highlighted in

red, respectively. These test images are specifically selected to represent diverse airports with variability in their structure in Fig.4(a) to Fig.4(f). As can be seen, the novel algorithm can locate airport regions successfully, even in the case of Fig.4(a) where the gray contrast between the background and the airport is very low. In particular, Fig.4(d) shows the result when the size changes greatly; in addition, the situation of revolution of plane has been taken into account that showed in Fig.4(e) and Fig.4(f). The situations of dark runway and being similar to surrounding environment are extremely rare in test data, our algorithm had poor effect in this case, but it will be our future work.

4.2 Comparison with Other Algorithms

Compared with traditional algorithm and long linear algorithm which use the line detection without line clustering, the proposed airport recognition algorithm can detect and recognize airport more accurately. We first provide a visual comparison among the airport recognition results from the other method and our proposed method in Fig.5(a). From the result images, the performance of traditional and long linear methods are not as good as ours. The other two methods both locate the airport incorrectly, even can't find the location of the airport. From a statistical point of view, we compute the overlap rate between these three methods and ground truth that is showed in Fig.5(b), and the overlap rate was defined in [5]. The x-coordinate of the broken line graph is overlap rate which changes from 0–100%; the greater the overlap rate is, the closer the result to the ground truth. And more, y-coordinate means numbers of results which is corresponding with overlap rate. Obviously, the red broken line focuses on the 80% overlap rate, respectively demonstrates the advantage of our method.

In the opinion of another statistical view, we obtain 90.48% TR and 14.41% FA rates which mean this result is very promising while traditional method have 36.99% TR and 33.42% FA rates, long linear method have 55.98% TR and 21.10% FA rates.

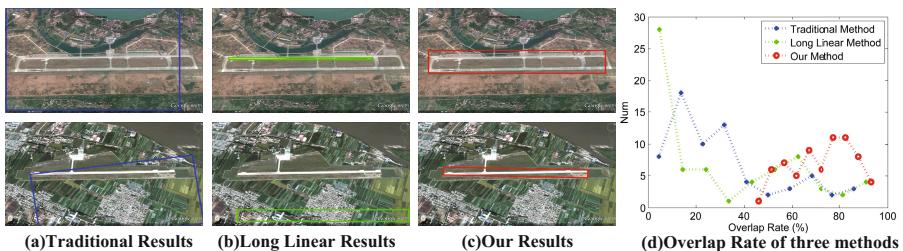


Fig. 5. Our algorithm is compared with other two algorithms mentioned in [4] and [7]. (a) shows the results of traditional algorithm[4], long linear algorithm[7] and our algorithm.(b) The comparison of Overlap rate about these three algorithms.

5 Conclusion

In this work, we proposed a novel local-global joint decision based clustering method, which contains a statistical probability model and a relative similarity model. For the sake of the inference of other straight lines as less as possible, the proposed approach locates airport regions more precise with the purpose of a better recognition effect. The main contribution of our paper is clustering all the straight lines into some subsets through the models of local-global joint decision based clustering automatically; the subsets which are equal of suspected regions may contain an airport in an attempt to save computation time. Afterward, these suspected regions are mapped to the normalized scale for airport recognition.

Compared with the other two traditional algorithms, our algorithm can locate an airport more accurately and distinguish the airport from the complex environment as the experimental results showed. Finally, it should be noted that the proposed local-global joint decision based clustering method has the potential to be extended to a group of other kind of targets in aerial images.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No.61231016, No.61272288), 2013 New People and New Directions Foundation of School of Computer Science in NPU (No.13GH014604), the NPU Foundation for Fundamental Research (No.JC201120, No.JC201148), and Plan of Soaring Star of Northwestern Polytechnical University (No.12GH0311).

References

1. Tao, C., Tan, Y., Cai, H., Tian, J.: Airport detection from large ikonos images using clustered sift keypoints and region information. *IEEE Geoscience and Remote Sensing Letters* 8(1), 128–132 (2011)
2. Pi, Y., Fan, L., Yang, X.: Airport detection and runway recognition in sar images. In: 2003 IEEE International Geoscience and Remote Sensing Symposium, vol. 6, pp. 4007–4009. IEEE (2003)
3. Liu, D., He, L., Carin, L.: Airport detection in large aerial optical imagery. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 5, pp. V–761. IEEE (2004)
4. Weng, W.: Research and implementation of airport and bridge technique for remote sensing image. Xidian University (2010)
5. Kalal, Z., Matas, J., Mikolajczyk, K.: Pn learning: Bootstrapping binary classifiers by structural constraints. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49–56. IEEE (2010)
6. Hao, Q., Ni, G., Guo, P., Chen, X., Tang, Y.: A method of automatic recognition of airport in complex environment from remote sensing image. In: International Conference on Optical Instrumentation and Technology, International Society for Optics and Photonics, pp. 751333–751333 (2009)

7. Cao, S., Jiang, J., Zhang, G., Yuan, Y.: Airport runway detection based on long linear structure. *Journal of Infrared and Laser Engineering* 41(4), 1078–1082 (2012)
8. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 24(7), 881–892 (2002)
9. Li, Y.: An adaptive k-means clustering algorithm. *Journal of Computer Research and Development* 2 (2007)
10. Yujian, L.: A clustering algorithm based on maximal θ -distant subtrees. *Pattern Recognition* 40(5), 1425–1431 (2007)

Ultra-Wideband Interference Suppression in Time Reversal Transmitted-Reference UWB System

Lan Zhang¹, Fang-Chao Zhang², and Bing Wang³

¹ School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing, China
Zhanglan2003@vip.sina.com

² School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing, China
15136220821@163.com

³ Department of Data Transmission,
Beijing Institute of Satellite Information Engineering, Beijing, China
wangbing0001@126.com

Abstract. Aiming to improve the performance of transmitted reference(TR) UWB system and mitigate external interference in multipath channel, time reversal(TiR) is employed. The BER performance of TR-UWB systems with and without TiR technology are investigated respectively, and the results imply that the SNR of TR-UWB system is improved about 10dB in multi-path channel with TiR pre-filter than that without TiR pre-filter whilst BER=1E-3. The impacts of the notch filter's parameters on performance of TiR TR-UWB system are investigated. The simulation results show that every increase of 528MHz in the width of notch filter causes 5dB degradation in the performance whilst the depth of notch filter has little influence on it. Finally, BER performance of this system with notch filter is studied in the case of MB-OFDM-UWB interference. Results indicate that with ultra-wideband notch filter technology, strong ultra-wideband interference within a bandwidth of $3 \times 528\text{MHz}$ can be mitigated. For BER=1E-2, the signal-to-interference ratio(SIR) of TR-UWB improves 25dB with a notch filter comparing to that without a notch filter. The present work is expected to be useful for the design of the simple and advanced TR-UWB system.

Keywords: time reversal, transmitted reference UWB, notch filter, interference suppression, MB-OFDM-UWB, BER.

1 Introduction

Ultra-wideband technology has gained significant interest after year 2002 when the US Federal Communications Commission(FCC) allocated frequency band from 3.1GHz to 10.6GHz for unlicensed operation of UWB radio and determined transmission power to be a maximum of under -43.1dBm/MHz[1]. Due to low transmission power and the high bandwidth, UWB communication systems suffer short communication distance and interferences from other communication systems. How to improve signal-to-noise ratio(SNR) to widen communication range under FCC power limit and how to coexist with other licensed or unlicensed communication systems are interesting topics.

In recent years, a signal focusing technique called time reversal(TiR) is often used in UWB systems to turn effect of multipath into benefit[2-4]. TiR is a transmission scheme in which impulse response of time reversal channel is used as a transmitter pre-filter. In dense multipath propagation channel, strong temporal compression can be achieved. A focusing gain can improve the transmission distance for communication purposes.

On the other hand, UWB communication systems are harmless to others communication systems because of their low transmission power, but they suffer the interference from other licensed or unlicensed communication systems. Recently, some works have focused on the coexistence issue among UWB systems. IR-UWB [5-8] and MB-OFDM-UWB[9-11] systems are the main approaches to UWB systems: The IR-UWB system is a candidate as a physical layer for low date rate wireless personal area network(WPAN); The MB-OFDM-UWB system is also a candidate as a physical layer for high data rate WPAN. IR-UWB and MB-OFDM-UWB systems have their different application areas but share the possibility of coexistence in the future.

Some related study results are presented in literatures. In literature [12], the authors use a multi-carrier template wave to mitigate the effect of MB-OFDM interference on a IR-UWB system. In literature [13], a waveforming technology is suggested to mitigate the interference caused by IR-UWB in MB-OFDM system. However, either adaptive multi-carrier template wave or a waveforming technology is difficult to carry out. Besides, limited researches touch the case when both receivers of UWB systems are integrated into one terminal, thus strong interference suppression technology requires further investigation.

In the MB-OFDM approach, the entire UWB spectrum is divided into 14 subbands, each has a bandwidth of 528MHz. MB-OFDM-UWB interference can be treated as narrowband interference(NBI) comparing to IR-UWB signal which occupies a pulse width of 100psec and corresponds to about 10GHz frequency bandwidth. Notch filtering technology is a mature approach to mitigate strong NBI. Some studies investigate the effects of notch filter on interference cancellation in IR-UWB systems [14-17]. In literature [14-15], the interference suppression with notch filters in IR-UWB systems are explored. In literature [16-17], the bit error rate(BER) performance of IR-UWB systems with notch filters are investigated in the presence of partial-band interference and wideband interference. These research results show notch filters are effective and simple in mitigating NBI, partial-band and wideband interference in IR-UWB systems. However, there are limited research focused on the impacts of parameters of notch filters on the performance of TR-UWB systems and the impacts of the notch filter on the performance of TiR IR-UWB systems. Based on the above statements, this research explores the application of TiR pre-filter in transmitted reference system(TR-UWB) and the impact of the notch filter on TR-UWB system with TiR pre-filter. The rest of this paper is organized as follows: the simulation model of TR-UWB system in multipath channel with TiR pre-filter is established; the performance of TR-UWB receivers with and without TiR pre-filter are investigated respectively; the impacts of the notch filter's parameters on the performance of TR-UWB are examined; the effect of notch filter suppressing MB-OFDM-UWB interference is studied.

2 Simulation Setup

2.1 System Diagram and Signal Models

The simulation diagram of TR-UWB Transceiver is shown in Figure 1. For easy illustration, this paper only focuses on a single-user case employing PPM modulation for data transmission. The transmitted signal $S_{RF}(t)$ in Figure 1 can be expressed as:

$$S_{RF}(t) = s(t) * h(-t) \quad (1)$$

$$s(t) = \sum_{j=-\infty}^{\infty} [p(t - jT_f - C_j^k t_c) + p(t - jT_f - C_j^k t_c - \zeta d_{[\frac{j}{N_s}]}^k - D)] \quad (2)$$

Where $s(t)$ is the signal of the TR-UWB, $h(-t)$ is channel impulse response of the time-reversal pre-filter, $p(t)$ is the shape of transmitted pulse, T_f is the duration time of a frame, C_j^k is the TH code of the k th user, t_c is the duration time of a chip, N_s is the number of pulses transmitted per symbol. ζ is the modulation index, $d_{[\frac{j}{N_s}]}^k$ is the binary data sequence of the k th user, D is the delay time of between modulated pulse and reference pulse.

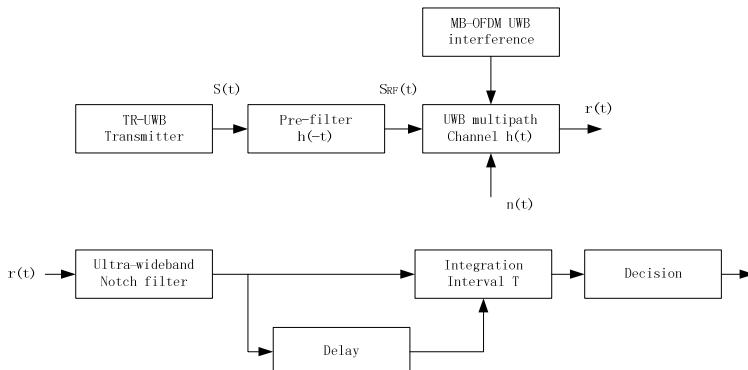


Fig. 1. Diagram of TR-UWB Transceiver

For interference signal of MB-OFDM-UWB, the transmitted sequence of the i th block can be written as:

$$s_i(k) = \sqrt{2P_T} \sum_{n=0}^{N-1} a_i(n) e^{\frac{j2\pi nk}{N}} \quad (3)$$

$$-G \leq k \leq N-1 \quad (4)$$

Where $a_i(n)$ is the i th symbol, N is the points of IDFT, $\sqrt{2P_T}$ is the transmit power, G is the guard samples, and $S_i(k)$ is assumed to be zero for $k < G$ and $K \geq N$, the total transmitted sequence at baseband can be written as:

$$s(k) = \sum_{j=-\infty}^{\infty} s_i(k - i(N + G)) \quad (5)$$

Then an RF carrier is inserted and the signal is taken to the specified carrier frequency with respect to the frequency-hopping pattern.

With the time reversal transmitter pre-filter, the received signal $r(t)$ in the absence of interference can be modeled as:

$$r(t) = [s(t) * h(-t) + n(t)] * h(t) = s(t) * R_{ss}(t) + n(t) * h(t) \quad (6)$$

Where $R_{ss}(t)$ represents the autocorrelation. The $R_{ss}(t)$ has a strong peak corresponding to the energy of all multipath components, thereby the SNR is improved significantly.

2.2 Channel Model and Simulation Conditions

The multipath UWB channel $h(t)$ is modeled as a single cluster of S-V channel as follow:

$$h(t) = \sum_{l=0}^{L-1} h_l \delta(t - t_l) \quad (7)$$

Where the total number of multipath is L , h_l represents the amplitude for l th path and t_l is denoted as arrival time for l th path. The TR-UWB system with the second derivative of Gaussian function is modulated in TH-PPM modulation. The MB-OFDM-UWB system operates in three sub-bands with center frequencies at 3432MHz, 3960MHz, and 4488MHz. The simulation parameters of multipath channel and signal are shown in Table 1.

Table 1. Parameters of multipath channel and signal

Multipath channel parameters($L=6$)		Simulation signal parameters	
Multipath Delaytime	Multipath Magnitude ($T_C=300$ psec)	TR-UWB parameters	MB-OFDM-UWB parameters
t_0 : 0psec	$h_0=1$	Pulse waveform: second derivative Gaussian	Operation frequency: Group1
t_1 : 150psec	$h_1=-h_0*e^{-(t_1/T_C)}$	Pulse interval: 5ns	Time frequency code: TFC1
t_2 : 300psec	$h_2=h_0*e^{-(t_2/T_C)}$	Repetitive number per bit: 12	FFT Points: 128
t_3 : 450psec	$h_3=-h_0*e^{-(t_3/T_C)}$	Pulse width: 100psec	Length of information: 242.42ns
t_4 : 600psec	$h_4=h_0*e^{-(t_4/T_C)}$	Delay: 2.5ns	Length of symbol: 312.5ns
t_5 : 750psec	$h_5=h_0*e^{-(t_5/T_C)}$		

The notch filter employs Chebyshev bandstop filter. Figure 2 shows the magnitude response of the notch filter with the notch width of 3×528 MHz, the notch depth of 50dB and a center frequency of 3.96GHz.

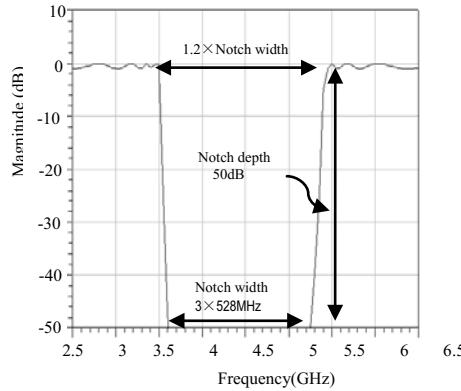


Fig. 2. Magnitude response of the notch filter in center frequency of 3.96GHz

The notch filter employs Chebyshev bandstop filter. Figure 2 shows the magnitude response of the notch filter with the notch width of 3×528 MHz, the notch depth of 50dB and a center frequency of 3.96GHz.

3 Simulation Results

3.1 Impulse Response of Channels with and without TiR Pre-Filter

The response of multipath channel single pulse with and without TiR pre-filter in the absence of noise and interference are shown in Figure 3. Figure 3(1) is the simulation

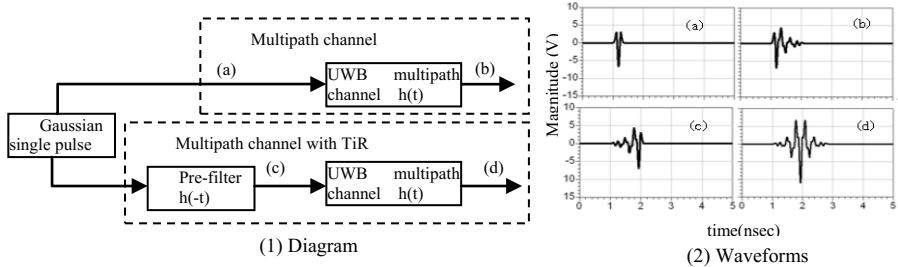


Fig. 3. Impulse response of channel: (1) Diagram (2) Waveforms

diagram while Figure 3(2) indicates its corresponding waveforms. A single second derivative of Gaussian pulse is employed. The waveforms of (a)~(d) in Figure 3(2) correspond to the second derivative of Gaussian pulse, impulse response of multipath channel, impulse response of TiR pre-filter, and impulse response of multipath channel with TiR pre-filter, respectively. Higher peak is observed in Figure 3(2)_d than in Figure 3(2)_b, indicating that SNR of TR-UWB system with the TiR technology is improved significantly.

3.2 BER versus SNR with and Without TiR Pre-Filter

Figure 4 shows the characteristics of BER versus SNR of TR-UWB systems in multipath channel with and without TiR pre-filter, respectively. In order to compare,

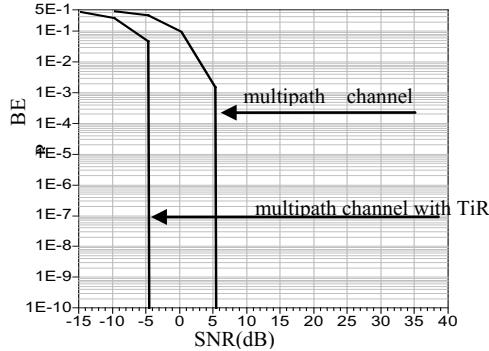


Fig. 4. BER versus SNR with and without TiR pre-filter

the transmission power are uniform. As we observed, The SNR has about 10dB improvement in multi-path channel with TiR pre-filter than that without in the case of BER=1E-3.

3.3 BER versus SNR with TiR for Various Parameters of Notch Filter

Figure 5 shows the characteristics of BER versus SNR with TiR for various widths of notch filters whilst the notch depth is 30dB. As we observed, as the width of notch filter increases from 0MHz to $3 \times 528\text{MHz}$ in a step of 528MHz, a standard sub-band

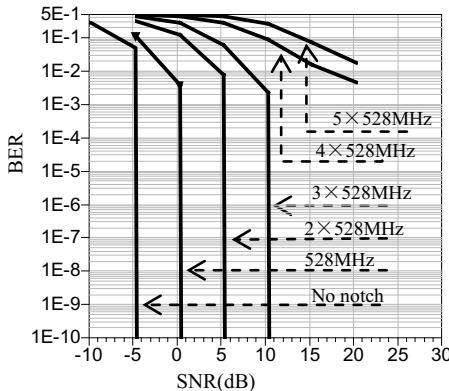


Fig. 5. BER versus SNR in the presence of 30dB depth of notch filter for various widths of notch filters

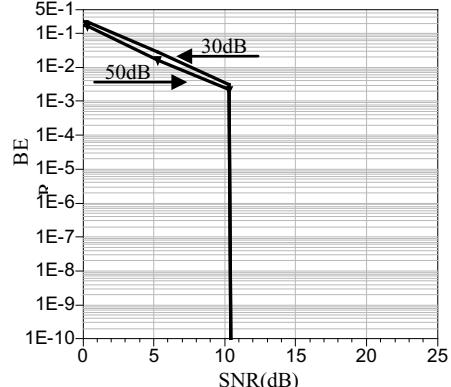


Fig. 6. BER versus SNR with TiR in the presence of 30dB and 50dB depth for $3 \times 528\text{MHz}$ width of notch filter

bandwidth of MB-OFDM-UWB system, the SNR degrades in a step of about 5dB in the case that BER=1E-3. When the width of notch filter is larger than $4 \times 528\text{MHz}$, the SNR illustrates significant degradation, for increasing the width of notch filter causes more loss of useful spectrum. The simulation results indicate that the width of notch filter is within $3 \times 528\text{MHz}$, TR-UWB systems can operate although with some extent's reduction of BER performance. Out of this range, the performance of system degrades considerably and may not be able to operate normally.

Figure 6 shows with the $3 \times 528\text{MHz}$ width of notch filter, the characteristics of BER versus SNR in the presence of 30dB and 50dB depth of notch filter respectively. No meaningful differences are observed between two curves implying that the depth of notch filter has little impact on the performance of system.

3.4 BER Performance in MB-OFDM-UWB Interference with and without Notch Filter in the Presence of TiR

Figure 7 shows the characteristics of BER versus signal-to-interference ratio(SIR) in the presence of MB-OFDM-UWB interference. The width of the notch filter is $3 \times 528\text{MHz}$ and the depth is 30dB. The solid curve indicates the characteristic of BER versus SIR with notch filter and the dotted curve without. The results indicate that when MB-OFDM-UWB operates in a bandwidth of $3 \times 528\text{MHz}$, the SIR of TR-UWB with a notch filter improves 25dB comparing to that without a notch filter in the case of BER=1E-2. The present research assumes that the application of wideband notch filter is a simple and effective way regarding the remove of a certain percentage of bandwidth of UWB interference.

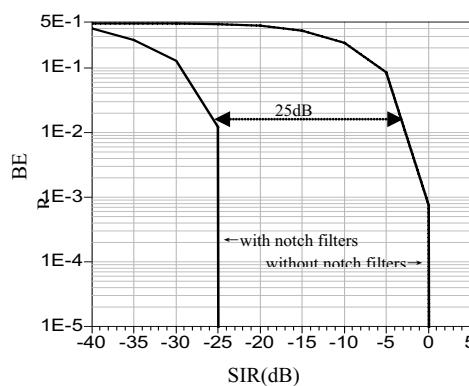


Fig. 7. BER versus SIR with and without notch filter

4 Conclusion

TiR and ultra-wideband notch filter technology are suggested and investigated in TR-UWB systems to improve performance and mitigate MB-OFDM ultra-band interference in this paper. The BER performance of TR-UWB system with and

without TiR technology are explored respectively, and the results imply BER performance shows about 10dB improvement with TiR technology. Second, the impacts of parameters of notch filters on the BER performance of TR-UWB system with TiR pre-filter are investigated, the simulation results indicate the performance suffers little influence from depth of notch filter, while it degrades with increasing notch filter's width. Within $3 \times 528\text{MHz}$ the width of notch filter, TR-UWB system can operate with BER performance degradation to some extent. Out of this range, performance of system degrades considerably and might not be able to operate normally. Finally, BER Performance in the presence of MB-OFDM-UWB interference with and without notch filter are studied. The results indicate with notch filter technology, strong ultra-wideband interference within a bandwidth of $3 \times 528\text{MHz}$ can be mitigated. The present work is expected to be useful for the design of simple and advanced TR-UWB systems.

References

1. Nekooga, F.: Ultra-Wideband Communications Fundamentals and Applications. Prentice Hall (August, 31, 2005) ISBN: 0-13-146326-8
2. Alizadeh, S., Khalegi Bizaki, H., Okhovvat, M.: Effect of channel estimation error on performance of time reversal-UWB communication system and itscompensation by pre-filter. *IET Communications* 6, 1781–1792 (2012)
3. Ishikawa, H., Matsumoto, A., Nakamura, R., et al: Time-Reversal UWB-IR Considering Channel Estimation Error. In: 2013 IEEE on Radio and Wireless Symposium (RWS), Texas, pp. 283–285 (2013)
4. Monsef, F., Cozza, A., Abboud, L.: Effectiveness of Time-Reversal technique for UWB wireless communications in standard indoor environments. In: 2010 Conference Proceedings, ICECom, Dubrovnik, pp. 1–4 (2010)
5. Hoctor, R.T., Tomlinson, H.W.: An overview of delay-hopped, transmitted-reference RF communications. *Technical Information Series* 2, 1–29 (2002)
6. Pardinas-Mir, J.A., Lamberti, R., Muller, M., Gimenes, C.: A fast low-cost TOA estimation for UWB impulse radio networks. In: 2012 9th International Conference on Communications (COMM), Bucharest, pp. 27–30 (2012)
7. Liang, Z., Dong, X., Jin, L., Gulliver, T.A.: Improved low-complexity transmitted reference pulse cluster for ultra-wideband communications. *IET Communications* 6(7), 694–701 (2012)
8. Fall, B., Elbahhar, F., Heddebaut, M., Rivenq, A.: Time-Reversal UWB positioning beacon for railway application. In: 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, NSW, pp. 1–8 (2012)
9. Liao, J.-C., Wu, Y.-W., Ma, H.-P.: Design of a high-spectral-efficiency MIMO MB-OFDM UWB baseband transceiver. In: 2010 IEEE 11th International Symposium on Spread Spectrum Techniques and Applications (ISITA), Taichung, pp. 151–154 (2010)
10. Lee, K.-M., Han, D.S.: Improved WiMedia system supporting MIMO for wireless HD STB and mobile device. In: 2011 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, pp. 447–448 (2011)
11. Le, N.-P., Tran, L.-C., Safaei, F.: Very high data rate MB-OFDM UWB systems with transmit diversity techniques. In: 2012 International Symposium on Communications and Information Technologies (ISCIT), Gold Coast, QLD, pp. 508–512 (2012)

12. Ohno, K., Ikegami, T.: Interference mitigation study for UWB radio using template waveform processing. *IEEE Transactions on Microwave Theory and Techniques* 54(4), 1782–1792 (2006)
13. Mehbodniya, A., Aissa, S.: Coexistence between DS-UWB and MB-OFDM : analysis and interference mitigation. In: *IEEE Global Communication Conference (Globecom 2007)*, Washington DC, pp. 5200–5204 (2007)
14. Wang, J., Tung, W.T.: Narrowband interference suppression in time-hopping impulse radio ultra-wideband communications. *IEEE Transactions on Communication* 54(6), 1057–1067 (2006)
15. Zhang, L., Wang, B., Zhang, Z.-F.: NBI suppression using double notch filtering in DS-UWB receivers. *High Technology Letters* 19(10), 1037–1040 (2009)
16. Cui, S., Teh, K.C., Li, K.H., Guan, Y.L., Law, C.L.: BER performance of transmitted-reference UWB systems with notch filter in the presence of inter-pulse interference and partial-band interference. In: *2007 6th International Conference on Information, Communications & Signal Processing*, Singapore, pp. 1–5 (2007)
17. Das, B., Das, S.: Interference mitigation techniques in transmitted reference UWB system used in WPANs NLOS channel environment. In: *2010 Annual IEEE India Conference (INDICON)*, Kolkata, pp. 1–4 (2010)

Exploiting the Wisdom of Crowd: A Multi-granularity Approach to Clustering Ensemble

Dong Huang¹, Jian-Huang Lai¹, and Chang-Dong Wang²

¹ School of Information Science and Technology, Sun Yat-sen University, China

² School of Mobile Information Engineering, Sun Yat-sen University, China

huangdonghere@gmail.com, stsljh@mail.sysu.edu.cn,

changdongwang@hotmail.com

Abstract. There are three levels of granularity in a clustering ensemble system, namely, base clusterings, clusters, and instances. In this paper, we propose a novel clustering ensemble approach which integrates information from different levels of granularity into a unified graph model. The normalized crowd agreement index (NCAI) is presented for estimating the quality of base clusterings in an unsupervised manner. The source aware connected triple (SACT) method is proposed for inter-cluster link analysis. By treating the clusters and the instances altogether as nodes, we formulate the ensemble of base clusterings and multiple levels of relationship among them into a bipartite graph. The final consensus clustering is obtained via an efficient graph partitioning algorithm. Experiments are conducted on four real-world datasets from UCI Machine Learning Repository. Experimental results demonstrate the effectiveness of our approach for solving the clustering ensemble problem.

1 Introduction

Data clustering is a fundamental problem in the field of data mining and knowledge discovery [1]. Its purpose is to partition unlabeled data into a certain number of groups or clusters. Different clustering methods or the same method with different parameters would lead to different clustering results for a dataset. Each method has its own merits and as well weaknesses, which is one of the main motivations for the clustering ensemble technique [2].

Clustering ensemble aims to combine multiple clusterings, each referred to as a base clustering or an ensemble member, to construct a so-called consensus clustering. In recent years the clustering ensemble technique has been drawing an increasing attention and many clustering ensemble approaches have been developed [2]. Among them, a popular category of approaches is based on the pairwise similarity matrix [3,4,5]. The approaches in this category first construct an $n \times n$ similarity matrix between n instances based on the information of multiple base clusterings. Then the hierarchical agglomerative clustering methods [1], such as single-link (SL) and complete-link (CL), can be performed on the similarity matrix and thus the consensus clustering is achieved. In [3] Fred and Jain proposed the evidence accumulation clustering (EAC) method, which computed a co-association matrix (or similarity matrix) S with the value of each entry obtained as $S(i,j) = m_{ij}/M$, where m_{ij} was the number of times that

instances i and j occurred in the same cluster among the M base clusterings. Wang et al. [5] proposed the probability accumulation method, which was a generalization to the EAC and took into consideration the cluster sizes of the base clusterings. Iam-On et al. [4] further exploited the relations between clusters and proposed two types of similarity matrices, namely, the connected-triple based similarity (CTS) and the Sim-Rank based similarity (SRS). These approaches fuse the information from base clusterings by a knowledge pool, the similarity matrix, to which each of the base clusterings makes contributions identically. However, there may be some better base clusterings as well as some worse ones, or even very bad ones in a clustering ensemble. How to evaluate and weight the base clusterings w.r.t. their quality remains an unaddressed problem.

An alternative category of clustering ensemble approaches is based on graph partitioning [6,7]. Strehl and Ghosh [6] formulated the ensemble of clusterings into a hypergraph, where the clusters could be represented as hyperedges. Three graph partitioning algorithms were further proposed in [6], that is, the cluster-based similarity partitioning algorithm (CSPA), the hypergraph-partitioning algorithm (HGPA), and the meta-clustering algorithm (MCLA). Fern and Brodley [7] introduced the hybrid bipartite graph formulation (HBGF), which treated the clusters and the instances as nodes in the same graph. A graph link between two nodes existed if and only if one node was an instance and the other was the cluster containing it. However, in the model of [7], the relations between clusters and between clusterings were not considered.

In this paper, we propose a multi-granularity clustering ensemble approach. An ensemble of base clusterings can be viewed as a crowd. And we exploit “the wisdom of the crowd” in three different levels, that is, between base clusterings, between clusters, and between clusters and instances. Compared with the existing approaches, the proposed approach is distinguished in three aspects. Firstly, we present the normalized crowd agreement index (NCAI) which evaluate the base clusterings w.r.t. the crowd of ensemble members. Secondly, a source aware connected triple (SACT) method is proposed for constructing linkage between clusters, which takes the common neighboring clusters and their reliability into consideration. Further, we model the clustering ensemble system into a bipartite graph, where both clusters and instances are treated as nodes and multiple levels of relationship are considered for graph link construction.

The remainder of this paper is organized as follows. The proposed clustering ensemble approach is introduced in Sect. 2. The experimental results are reported in Sect. 3. We conclude this paper in Sect. 4.

2 Proposed Clustering Ensemble Framework

2.1 Problem Formulation

Given a dataset $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, where x_i is the i -th instances and n is the cardinality of \mathcal{X} . Consider M base clusterings (or partitions) of the dataset \mathcal{X} , and let

\mathcal{P} denote the set of the M base clusterings. Then the clustering ensemble is defined as

$$\mathcal{P} = \{P^1, P^2, \dots, P^M\}, \quad (1)$$

$$P^1 = \{C_1^1, C_2^1, \dots, C_{n_1}^1\}, \quad (2)$$

⋮

$$P^M = \{C_1^M, C_2^M, \dots, C_{n_M}^M\} \quad (3)$$

where P^i is the i -th base clustering in \mathcal{P} , C_j^i is the j -th cluster in P^i , and n_i is the number of clusters in P^i . For $i = 1, \dots, M$, it holds that $\cup_{j=1}^{n_i} C_j^i = \mathcal{X}$. The set of clusters from all base clusterings is denoted as $\mathcal{C} = \{C_1, \dots, C_{n^c}\}$, where $n^c = \sum_{i=1}^M n_i$.

The problem is to find a consensus clustering P^* by summarizing the information from the clustering ensemble \mathcal{P} . This paper addresses the clustering ensemble problem in a hybrid manner. The relations in different levels of granularity are incorporated and a bipartite graph model is constructed with both clusters and instances treated as nodes. In the following, we will discuss the issues of base clustering evaluation, inter-cluster link analysis, and the bipartite graph model respectively in Sect. 2.2, 2.3, and 2.4.

2.2 Crowd Agreement Estimation

Due to the diversity of clustering algorithms and datasets, there may be good-quality base clusterings as well as inferior ones, or even very bad ones, in a clustering ensemble. Compared to the inferior base clusterings, the good ones are more likely to contribute positively to the clustering ensemble system. There is a need to give the good base clusterings a bigger say while lessening the influence of the bad ones to a certain extent. The key issue here is to evaluate the base clusterings without knowing the ground-truth.

Some methods have been developed to estimate the clustering quality by utilizing the within-cluster and between-cluster information [8,9]. These methods are only applicable to numerical data and need access to the feature vectors of the data instances. But for the clustering ensemble problem, the given information is the labeling of multiple clusterings (as shown in (1), (2), and (3)), while the feature vectors of the data are generally not supposed to be given. Rather than exploring the distribution of the data features, we estimate the quality of each base clustering by consulting the other individuals in the ensemble.

In social and economic science, “the wisdom of the crowd” is the process of taking into consideration the opinion of a crowd of individuals rather than a single expert [10]. In our work, to estimate the quality of a base clustering without supervision, we collect information from the crowd of base clusterings. Each base clustering is compared with the other ones in the ensemble and the average opinion is obtained for quality estimation. We define the crowd agreement index (CAI) for base clustering P^i as follows:

$$CAI(P^i) = \frac{1}{M-1} \sum_{P^j \in \mathcal{P}, i \neq j} Sim(P^i, P^j), \quad (4)$$

where $Sim(P^i, P^j)$ represents the similarity between the two base clusterings P^i and P^j . The base clustering that gains the maximum agreement from the crowd is treated as

a reference object, and the reliability of other members is estimated by comparing their crowd agreement to the maximum agreement. Then we define the normalized crowd agreement index (NCAI) as follows:

$$NCAI(P^i) = \frac{CAI(P^i)}{\max_{P^i \in \mathcal{P}} CAI(P^i)}. \quad (5)$$

As it is defined, for $i = 1, \dots, M$, it holds that $NCAI(P^i) \in [0, 1]$. In this paper, the normalized mutual information (NMI) [6] is used as the similarity measure $Sim(P^i, P^j)$. The main idea here is to exploit the collective opinion from a crowd of diverse individuals. Other similarity measures or voting methods can also be utilized. The greater the NCAI of a base clustering is, the better its quality is supposed to be.

2.3 Inter-Cluster Link Analysis

This section investigates the relations among clusters and introduces the source-aware connected triple (SACT) method which measures the similarity of two clusters w.r.t. their commonly neighboring clusters and the source reliability.

Two clusters C_i and C_j are defined to be neighboring if they share some common instances, i.e., $C_i \cap C_j \neq \emptyset$. The Jaccard similarity is often used to measure the similarity between two sets (or clusters), which is defined as $J(C_i, C_j) = |C_i \cap C_j| / |C_i \cup C_j|$. The sharing instances of two clusters are taken in consideration in the Jaccard similarity. However, to estimate the relationship between two clusters in a clustering ensemble, the information from the *surrounding* clusters could also be exploited. Iam-On et al. [4] measured the similarity between clusters w.r.t. their common neighbors. But the reliability of these neighbors were not taken into consideration in [4].

With each base clustering treated as a source of clusters, the overall reliability of the clusters in a base clustering is correlated to the quality of that base clustering. In this work, we propose the source-aware connected triple (SACT) method, which extends the connected-triple based similarity (CTS) [4] and collects opinion from the neighboring clusters w.r.t. the reliability of the source. Formally, the SACT term between two clusters C_i and C_j w.r.t. a cluster C_k is defined as follows:

$$SACT_{ij}^k = I_{NCAI}(P(C_k)) \cdot \min(J(C_i, C_k), J(C_j, C_k)), \quad (6)$$

where $P(C_k)$ represents the base clusterings that C_k belongs to, and

$$I_{NCAI}(P^l) = (NCAI(P^l))^{\beta} \quad (7)$$

is the influence of the NCAI of P^l . The $\beta > 0$ is a parameter for adjusting the influence. For $l = 1, \dots, M$, it holds that $I(P^l) \in [0, 1]$. Following that, the SACT between C_i and C_j w.r.t. the clustering ensemble is defined as follows:

$$SACT_{ij} = \sum_{C_k \in \mathcal{C}} SACT_{ij}^k. \quad (8)$$

Then we have the similarity between C_i and C_j as

$$SIM_{SACT}(C_i, C_j) = \begin{cases} 1, & \text{if } i = j, \\ \frac{SACT_{ij}}{\max_{\forall C_x, C_y \in \mathcal{C}} SACT_{xy}}, & \text{otherwise.} \end{cases} \quad (9)$$

2.4 Bipartite Graph Construction and Partitioning

This section models the clustering ensemble problem as a bipartite graph partitioning problem. Both the clusters and instances are mapped onto the graph nodes. A graph link is constructed between two clusters, or between an instance and the cluster containing it. To implement a bipartite structure, each cluster in \mathcal{C} is used *twice*, which means each cluster is correlated to two nodes that lie in the two parts of the bipartite graph respectively. This will not be ambiguous for obtaining the final clustering result, as the final consensus clustering is achieved by *only* considering the instance nodes in each of the segments produced by graph partitioning.

Formally, the bipartite graph is defined as $G = (U, V, L)$, where $U = \mathcal{X} \cup \mathcal{C}$ and $V = \mathcal{C}$ are the nodes and L is the set of links. There are no graph links between the nodes in U or between the nodes in V . The link between an instance and the cluster containing it is weighted w.r.t. the NCAI, while that between two clusters is weighted w.r.t. the SACT. For two nodes $u_i \in U$ and $v_j \in V$, the weight of the link between them is defined as

$$w_{ij} = \begin{cases} \alpha \cdot I_{NCAI}(P(v_j)), & \text{if } u_i \in \mathcal{X}, v_j \in \mathcal{C}, u_i \neq v_j, \\ SIM_{SACT}(u_i, v_j), & \text{if } u_i \in \mathcal{C}, v_j \in \mathcal{C}, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

where $\alpha > 0$ is a parameter. Then we utilize the Tcut algorithm [11] to the bipartite graph G and a set of segments are produced. The data instances in each of these segments form a final cluster. Therefore the final consensus clustering for the clustering ensemble is achieved. Theoretically it is possible that there are no instance nodes in some segment and thus the actual cluster number would be less than we specified. However, this situation has not been encountered in our experiments probably due to that the links between a cluster and the many instances inside it are generally strong enough to hold at least part of them together.

3 Experimental Results

This section presents an empirical evaluation of the proposed clustering ensemble approach. We apply the proposed approach to several real-world datasets and compare its performance with the other clustering ensemble approaches.

3.1 Set Up

Four real-world datasets from the UCI Machine Learning Repository [12] are used, namely, *Seeds*, *Yeast*, *Pen-Digits*, and *Letter-Recognition*. The details of the datasets are given in Table 1. In this paper, the two parameters α and β of the proposed method are set to 0.5 and 2 respectively for the experiments on *every* benchmark dataset.

The proposed approach makes no specific assumptions about the ensemble of base clusterings. However, as it is the often case for ensemble systems, diversity is preferred. For generating the base clusterings, five different methods are used, namely, k -means, spectral clustering (SPEC), rival penalized competitive learning (RPCL) [13], affinity propagation (AP) [14], and hierarchical mode association clustering (HMAC) [15].

Table 1. Description of the benchmark datasets

Dataset	<i>Seeds</i>	<i>Yeast</i>	<i>Pen-Digits</i>	<i>Letter-Recognition</i>
#Instance	210	1,484	10,992	20,000
#Class	3	10	10	26
Dimension	7	8	16	16

3.2 Experimental Evaluation and Analysis

Comparison with Base Clusterings. We apply the proposed method to each benchmark datasets and construct the consensus clustering based on the base clusterings produced by the five methods listed in Sect. 3.1. The NMI measures of the base clusterings and the consensus clusterings are shown in Table 2. The consensus clustering constructed by our method outperforms every base clusterings for each dataset. A high-quality consensus clustering is obtained for the *Pen-Digits* dataset, whose NMI score reaches 0.809. Though the base clustering produced by SPEC for the *Seeds* dataset and the one produced by AP for the *Letter-Recognition* are both very low-quality, the NMI scores being 0.042 and 0.120 respectively, the proposed approach is still able to construct a much better consensus clustering with an NMI score around 0.5. The consensus clustering for the *Yeast* dataset only achieves an NMI of 0.233, which is mainly due to the overall poor performance of the ensemble of base clusterings. Still, the consensus clustering by our approach outperforms every base clusterings for this dataset.

Comparison with other Clustering Ensemble Methods. Here, we compare the consensus clusterings produced by our methods (with and without NCAI) and the other ten clustering ensemble methods, including the hybrid bipartite graph formulation (HBGF) [7] and nine similarity matrix based methods. The evidence accumulation clustering (EAC) [3], connected-triple based similarity (CTS) [4], and the SimRank based similarity (SRS) [4], are used for computing the similarity matrices, to each of which three agglomerative methods [1], namely, single-link (SL), complete-link (CL), and average-link (AL), are applied respectively for obtaining the final consensus clusterings. Experimental results of our method with and without NCAI are also evaluated. For each

Table 2. Comparing the proposed method with the base clusterings in terms of NMI

	<i>Seeds</i>	<i>Yeast</i>	<i>Pen-Digits</i>	<i>Letter-Recognition</i>
Our Method	0.578	0.233	0.809	0.492
<i>k</i> -means	0.297	0.178	0.539	0.450
SPEC	0.042	0.026	0.289	0.360
RPCL	0.292	0.179	0.542	0.445
AP	0.411	0.119	0.467	0.120
HMAC	0.409	0.059	0.726	0.465

Table 3. Comparison of different clustering ensemble methods in terms of NMI

Method	Seeds		Yeast		Pen-Digits		Letter-Recognition	
	Best- k	True- k	Best- k	True- k	Best- k	True- k	Best- k	True- k
Our Method (with NCAI)	0.578	0.466	0.233	0.233	0.809	0.809	0.492	0.396
Our Method (without NCAI)	0.528	0.468	0.175	0.144	0.782	0.768	0.485	0.383
HBGF	0.539	0.465	0.191	0.156	0.785	0.656	0.488	0.372
EAC+SL	0.466	0.045	0.094	0.033	0.334	0.009	0.257	0.075
EAC+CL	0.376	0.170	0.171	0.051	0.631	0.402	0.419	0.152
EAC+AL	0.507	0.473	0.174	0.066	0.790	0.706	0.485	0.335
CTS+SL	0.540	0.036	0.100	0.037	0.340	0.003	0.261	0.100
CTS+CL	0.425	0.260	0.178	0.084	0.723	0.589	0.460	0.294
CTS+AL	0.525	0.478	0.183	0.100	0.777	0.583	0.496	0.372
SRS+SL	0.518	0.034	0.090	0.035	0.491	0.001	0.265	0.066
SRS+CL	0.413	0.268	0.159	0.122	0.761	0.739	0.436	0.352
SRS+AL	0.469	0.422	0.170	0.092	0.789	0.685	0.484	0.339

method, the number of clusters k for the consensus clustering is set to two values respectively, that is, True- k and Best- k . True- k is the number of true classes in a dataset. Best- k is the number of clusters that lead to the optimal performance for the dataset.

As shown in Table 3, much better results for the *Yeast* dataset and better results for *Pen-Digits* and *Letter-Recognition* are obtained by our method with NCAI than that without NCAI. For the *Seeds* dataset, the consensus clusterings by our method (without NCAI), the EAC+AL, and the CTS+AL for True- k are slightly better than that by our method with NCAI. But the with-NCAI version of our method significantly outperforms the without-NCAI version of ours, the EAC+AL, and the CTS+AL for Best- k on *Seeds*. For *Yeast*, *Pen-Digits*, and *Letter-Recognition*, the results by our method, in terms of both Best- k and True- k , are among the best in Table 3. Comparing with the baseline clustering ensemble methods, our method achieves a better overall performance.

Choices of Parameters. There are two parameters in our approach. Parameter α is the scale factor for the link weight between an instance and the cluster containing it. Parameter β adjusts the influence of the NCAI. As can be seen in Table 4, the proposed approach is very stable w.r.t. the two parameters. Empirically, it is suggested that α be set in the interval of $(0.1, 1)$ and β in $(1, 4)$ for different datasets.

Table 4. The performance of our approach with varying parameters in terms of NMI

α	0.5					0.01	0.1	1
β	0	1	2	4	8	2	2	2
<i>Seeds</i>	0.528	0.561	0.578	0.709	0.595	0.553	0.539	0.578
<i>Yeast</i>	0.168	0.212	0.233	0.246	0.243	0.220	0.237	0.235
<i>Pen-Digits</i>	0.782	0.789	0.809	0.803	0.793	0.434	0.780	0.803
<i>Letter-Recognition</i>	0.485	0.493	0.492	0.507	0.501	0.302	0.496	0.496

4 Conclusion

This paper proposes a novel clustering ensemble approach which formulates the cumulative information from different levels of granularity into a unified graph model. With the ensemble of base clusterings viewed as a crowd, the quality of each individual is estimated via collecting opinion from the other individuals in an unsupervised manner. The normalized crowd agreement index (NCAI) is present to measure the reliability of each base clustering. For analyzing the linkage between two clusters, we take the common neighboring clusters and the source reliability into consideration and propose the source aware connected triple (SACT) method. The clustering ensemble is further modeled into a bipartite graph by treating both clusters and instances as nodes and constructing graph links w.r.t multiple levels of relationship. The final consensus clustering is obtained by partitioning the graph. We conducted experiments on four real-world datasets from UCI Repository. Experimental results show that our method achieves better performance as compared with the baseline clustering ensemble methods.

Acknowledgements. This work was supported by NSFC (61173084 and 61128009), National Science & Technology Pillar Program (No. 2012BAK16B06).

References

1. Jain, A.K.: Data clustering: 50 years beyond k -means. *PRL* 31(8), 651–666 (2010)
2. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. *IJPRAI* 25(3), 337–372 (2011)
3. Fred, A.L.N., Jain, A.K.: Combining multiple clusterings using evidence accumulation. *TPAMI* 27(6), 835–850 (2005)
4. Iam-On, N., Boongoen, T., Garrett, S.: Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In: *ICDS* (2008)
5. Wang, X., Yang, C., Zhou, J.: Clustering aggregation by probability accumulation. *Pattern Recognition* 42(5), 668–675 (2009)
6. Strehl, A., Ghosh, J.: Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *JMLR* 3, 583–617 (2002)
7. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: *ICML* (2004)
8. Wu, S., Chow, T.W.S.: Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition* 37(2), 175–188 (2004)
9. Vendramin, L., Campello, R.J.G.B., Hruschka, E.R.: On the comparison of relative clustering validity criteria. In: *SDM* (2009)
10. Surowiecki, J.: *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Anchor Books (2004)
11. Li, Z., Wu, X.M., Chang, S.F.: Segmentation using superpixels: A bipartite graph partitioning approach. In: *CVPR* (2012)
12. Bache, K., Lichman, M.: UCI machine learning repository (2013)
13. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. *TNN* 4(4), 636–649 (1993)
14. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* 315, 972–976 (2007)
15. Li, J., Ray, S., Lindsay, B.G.: A nonparametric statistical approach to clustering via mode identification. *JMLR* 8(8), 1687–1723 (2007)

Spatio-temporal Features for Efficient Video Copy Detection

Ruijuan Hu¹, Bing Li², Weiming Hu², and Jinfeng Yang¹

¹ College of Aviation Automation, Civil Aviation University of China, Tianjin, China
`rjhu0525@163.com, jfyang@cauc.edu.cn`

² Institute of Automation, Chinese Academy of Sciences, Beijing, China
`{bli,wmhu}@nlpr.ia.ac.cn`

Abstract. Content-Based Video Copy Detection (CBVCD) aims at detecting whether or not a query video is a copy or part of a reference video from database. In this paper, we present a CBVCD system based on spatio-temporal features that can competitively deal with large database in terms of both performance and efficiency. Instead of selecting keyframes or uniformly sampling from original videos and then extracting global or local visual features for frames, we first divide a video into segments with fixed length and then extract 3D spatio-temporal features for the whole segment. After that, we perform similarity search comparing all the reference segments with query segments and apply a copy verifying to decide the final copy detection result. The experimental results on the TRECVID 2011 video copy detection dataset show that the proposed system is effective and efficient.

Keywords: Content-Based video copy detection, spatio-temporal features, similarity search, copy verifying.

1 Introduction

The goal of Content-Based Video Copy Detection (CBVCD) is to locate video fragments within a query video that are copies of reference videos. It is essential for many applications, for example, illegal content monitoring, copyright control, tracking the source and so on.

1.1 Related Work

In CBVCD, the copied videos are usually subject to various tolerated transformations (TTs) such as camcording, picture-in-picture(PIP), strong re-encoding, frame dropping, cropping, stretching, contrast changing, etc [1]. Some of these transformations are intrinsic to the video creation process, others are introduced intentionally for specific use. The transformation applied to a video can be one of the TTs mentioned above or combination of some of them. Besides, a query video can also be compiled in three modes: 1) only keep the reference video segment; 2) only keep the non-reference video segment; 3) inserting the reference

video segment into the non-reference video segment at a random offset. In addition, in any video task, the dataset is always much larger than image task. So the cost of computation and the detection efficiency must be considered at the same time. All the aforementioned problems make the task more challenging.

Most existing CBVCD systems are based on visual cues, which can be roughly divided into two categories: frame-based and video-based. Frame-based methods typically extract 2D interest points on selected keyframes or uniformly sampled frames of the videos and then use local descriptors to represent them [2]. These descriptors indicate spatial information significantly, while neglect temporal information. In order to be more discriminative for video task, temporal information is introduced via post-processing. Douze et al. [3] report their system which depends on bag-of-features combined with Hamming embedding of the frames. Then they determine the time shift using 1D Hough voting algorithm, and a 2D affine transformation is estimated between temporally consistent frame matches. In [4], R.Cameron first extracts SURF features [5] for the frames and then creates temporal signature by sorting the SURF feature counts in each region along the time-line. Although those frame-based algorithms are spatio-temporal to some extent and have achieved significant result, they have some obvious limitations. One is that they largely depend on the selection of frames. For uniformly sampled frames, there is no guarantee that the same frames will be selected both in reference and query videos unless for the assumption that the scene changes slowly so adjacent frames are similar. And the data is usually large. For keyframes, though the number of frames is much less, the system is highly lied on robustness of the shot-boundary detection. Moreover, in these frame-based method, the spatial and temporal information is not processed at the same time, it is hard to guarantee the correspondence of spatio-temporal information. For video-based approaches, trajectories are proposed by means of tracking 2D interest points throughout the video sequence. Law et al. [6] use 2D Harris detector and Kanade-Lucas-Tomasi (KLT) feature tracking for CBVCD. Although the local descriptor is enhanced with temporal information by using trajectories and have achieved promising result, the redundancy of the local descriptor is reduced. Moreover, it adds additional computations due to the need for tracking interest points over the whole video frames.

1.2 Our Work

Considering the limitations discussed above, we propose an alternative video copy detection system based on spatio-temporal features [7]. Fig. 1 gives an overview of our system. Having preprocessed videos, we first cut them into segments (a set of consecutive frames) and then extract spatio-temporal interest points from segments instead of spatial keypoints from frames. Extended scale-invariant feature transform (SIFT) features combined with PCA algorithm are extracted to represent these segments [8]. By comparing segments similarity and copy verifying, we can obtain final detection result.

The most important improvement of our method is that we directly use 3D interest points. Different from an image $I(x, y)$, here we must operate interest

point detector on a stack of images denoted by $I(x, y, t)$, making localization proceed not only along the spatial dimensions x and y but also the tempoal dimension t . The third dimension sufficiently represents the sequence of frames, which is the fundamental difference from images. In this regard, our spatio-temporal features based system has several advantages:

- The interest points are detected not only spatially but also over time, which makes them more discriminative as well as better localized within the segments.
- The resulting set of descriptors contains information from the whole segment, rather than focusing on a few selected frames.
- We extract features both spatially and tempoally at the same time, so there is no post-processing for adding the temporal cues.
- The PCA-SIFT descriptor, which is more discriminative and has lower dimension than SIFT, makes our system much more efficient.

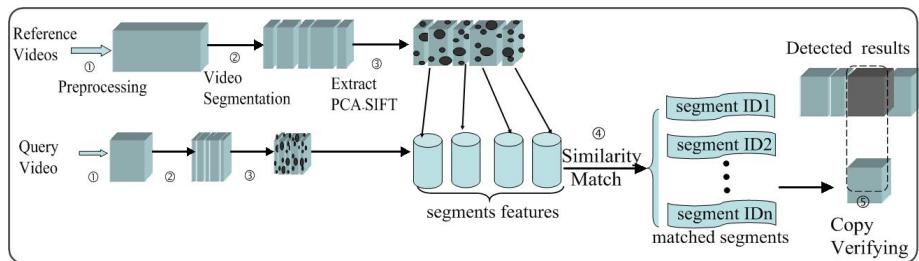


Fig. 1. An overview of the proposed system

2 Proposed System

In this section, we introduce our system based on spatio-temporal features in detail. For convenience, let R be the set of reference videos, let Q be the set of query videos. As shown in Fig. 1, the proposed system involves five parts in total: (1) Preprocessing. This step is to process input videos to diminish the effect of TTs. This creates a new set of reference videos R' and query videos Q' ; (2) Video Segmentation. This part partitions videos into short segments with fixed length; (3) Spatio-Temporal Feature Extraction. This step detects spatio-temporal interest points and represents them with PCA-SIFT; (4) Segment Similarity Match. It performs k-nearest neighbor method (KNN) for each query segment and returns most similar reference segments; (5) Copy Verifying. This part compares all the candidates for each query segment, and returns the final detection result.

2.1 Preprocessing

For better detection results, we preprocess input videos with two procedures: (1) skip frames that contain little information (2) diminish transformations effect.

For skipping frames, this task detects black frames by computing the variance between the intensity of frame pixels, and those frames whose values are under

a threshold will be skipped. We also skip exceptional frames, which is an outlier compared with its former frame f_{i-1} and the next f_{i+1} [9].

To diminish TTs effect, we focus on camcording and PIP that are very difficult to detect in CBVCD. The method is to detect persistent strong lines using Hough lines, which are consistent over the whole video, thus referring to the boundary of camcording or the window of PIP. If the edge lines are not vertical or horizontal, we use them to form a wrapping quadrilateral, the biggest one is seen as camcording boundary, then a new query is created by mapping the detected quadrilateral to the video corners. If most of the detected lines are vertical or horizontal, we remove short edge lines according to the size of PIP window (one third to half of the original video size) and merge others into a regular rectangle, we then build two new query videos, one is the foreground, another is the background. As we cannot guarantee the preprocessing to be completely correct, we process both the pre-processed and original query video with this system and determine the final result by choosing the more similar one.

2.2 Video Segmentation

In order to extract spatio-temporal features, we need partition every video into short segments. Here we do not divide a video into shots based on boundary because this will result in too few segments and largely depend on the efficiency of the boundary detection algorithm. In our system, each video is partitioned into segments with fixed length of 25 frames (less than a second, as the fps is 30 frames/s). The reason is that if the frames within a segment are too few, content of these frames are almost identical, the detected interest points in latter step are too sparse. And that if the frames are too many, the length of extracted features in a single segment is very high, which is a great challenge for efficiency and storage. To make a balance choice, we choose 25 in a segment to be discriminative enough as well as limited computation costs.

2.3 Spatio-Temporal Feature Extraction

After obtaining video segments, this section is to extract the spatio-temporal features. Different from most methods that extract spatial descriptors and add temporal information in the post-processing step, we directly extract features containing both spatial and temporal information, known as spatio-temporal features that are widely used in behavior recognition. The general idea of feature extraction is similar to the spatial case. First, We need a response function and find the interest points where the function reaches its local maxima. This step considers both the spatial and temporal information since the response function has two parameters σ and τ , corresponding roughly to the spatial and temporal scale of the detector. Then, at each interest point, a cuboid is extracted. To further represent the cuboid, in this paper we use the flattened gradient as the descriptor, which is essentially a generalization of the PCA-SIFT descriptor.

Interest Points Detection. The most widely used spatio-temporal interest point operator is proposed by Laptev and Lindeberg [10] that extends the 2D scale-invariant Harris-Laplace corner detector into the spatio-temporal domain. The basic idea is to find a spatial corner in an image region whose velocity vector is reversing direction. This Harris detector has proved to be efficient in many applications. Nevertheless, for some videos in our dataset, the true spatio-temporal corners are quite rare, greatly affecting the detection result. So, here we use an alternative detector proposed by P.Dollar in [7] whose feature set is more dense than the Harris detector. First, we define the response function as

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2. \quad (1)$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, defined as $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$, $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. Usually, let $\omega = 4/\tau$, so R simply correspond to σ and τ .

As h_{ev} and h_{od} are periodic, variations in local image intensities that contain frequency components will evoke the strongest response. This property is very important in behavior recognition, like people waving or bird flapping its wings. In our dataset, periodic contents are not so common. But this response function is still available because it also responds strongly to spatio-temporal corners. Areas undergoing drastic changes along temporal dimension or with spatially distinguishing features can induce strong response. Due to this ability, it is good at detecting 3D interest points, where R reaches local maxima.

Cuboids and Descriptor. At each interest point, a cuboid is extracted. This cuboid contains most of the volume of data that contribute to the response function. Given a large number of cuboids in our video dataset, we use a descriptor to represent each cuboid which can be computed once off-line. This descriptor is required to be discriminative and invariant to most of the transformations. The simplest way is to create a vector of flattened cuboid values by computing the gradient or Lucas-Kanade optical flow [11] of that cuboid. As the optical flow is more often used to extract motion information, which is not common in our videos, we focus more on the gradient. It has been proved that SIFT is effective in various video retrieving and near duplicate image detection task. So in this work, we adopt an extension of Lowes SIFT [12]. A cuboid is divided into regions and the extended SIFT is created by sampling the magnitudes and orientations of 3 axis-aligned gradient in that cuboid around the interest point. Then smoothed local orientation histograms are built which capture the important aspects of that cuboid, creating a high-dimension features. Then we use PCA to reduce the dimensionality of these descriptors. This idea is from Yan Ke [8], known as PCA-SIFT. The whole procedure can be summarized in the following steps: (1) given a segmented video, extract all the cuboids of segments set (2) compute descriptors for each cuboid (3) create an eigenspace by computing the covariance matrix of these vectors, and the top m eigenvectors are used as the projection matrix for PCA-SIFT (4) project all the descriptor vectors

using the eigenspace and result in new descriptor. This effectively linearly-project high-dimension vectors onto a low-dimensional feature space.

2.4 Segment Similarity Match

This task is to compute distance between two descriptors to determine whether the two vectors belong to the same cuboid in different segments. Distance between the descriptors can be calculated by using Euclidean. Then we perform KNN to retrieve the most similar reference segments for every query segment and obtain the k closest reference segments. Note that if the input data is large, we need to build an effective index to improve the search efficiency, such as vocabulary tree combined with inverted file [13].

2.5 Copy Verifying

The objective of the last step is to compare the candidate segments for each query segment and determine the final detection result. We opt to use an aggregate votes algorithm. In order to improve the detection accuracy, the votes $S_f(v)$ for any reference video v ($v \in R'$) are combined with a weighted value

$$S_f(v) = \sum_{i=1}^m \sum_{j=1}^n \omega_i^j S(s_i, r_i^j). \quad (2)$$

where ω_i^j is the weighted value and $\omega_i^j \in (0, 1]$, $S(s_i, r_i^j)$ is the similarity score of query segment s_i and reference segment r_i^j , r_i^j is a segment of v . Obviously, we can use distance value to replace this similarity score, but to be simple, we normalize it as $S(s_i, r_i^j) = 1$. This task is calculated as follows:

- (1) For query segment , add rank information $rank_i^j$ to the similarity list.
- (2) Compute the corresponding w_i , as $w_i = 1 - (rank_i^j - 1) * (1/k)$.
- (3) Compute votes for all reference videos and aggregate the votes, then locate the maximum value and if the maximum is larger than a threshold, then the copy detection result is , and if it is less than the threshold, there is no copy.

3 Experiments

In this section, we evaluate our system on TRECVID 2011 dataset [1]. This dataset contains more than 12,000 reference videos and over 10,000 query videos which are created with TTs. Our system is tested on a subset of TRECVID 2011 dataset and perform two experiments to show the results.

3.1 Effectiveness of Preprocessing

To assess the impact of preprocessing on features match, a simple but typical image similarity detection experiment is performed and the results are listed in Table 1. All the images used in this experiment are frames of videos from TRECVID 2011 dataset, containing about 60,000 reference images and 4000

query images. We extract SIFT features for each frame and bag-of-words is used combined with inverted files [13]. In the last step we compare distance using Euclidean distance and return 30 most similar frames for each query frame.

We can clearly find that camcording and PIP are difficult (only 61% or so) to be correctly detected compared with other TTs which can achieve an accuracy of nearly 84%. However, even the preprocessing is not perfect enough, it indeed helps to improve the feature matches, making it an essential step in our system.

Table 1. Accuracy of preprocessing, image similarity detection

	Preprocessing	Image similarity detection (no preprocessing)	Image similarity detection (with preprocessing)
camcording	78.6%	61.9%	77.4%
PIP	84.7%	60.6%	80.7%
T3,T4,T5,T6,T8,T10	—	83.3%	—

3.2 Effectiveness of Spatio-temporal Feature-Based System

In this experiment, after preprocessing the input videos, we segment the videos with a fixed length of 25 frames. After extracting the SIFT, we set the PCA coefficient number $m = 200$. In the Segment Similarity Match step, we set $k = 20$ meaning we select the top 20 closest segments for each query segment. To make a comparison we also perform an frame-based video copy detection system (with preprocessed) with SIFT extraction and to speed up the procedure we apply vocabulary tree combined with inverted file [13]. In order to evaluate our systems performance, we measure recall, precision and F1. For evaluating the computation cost, we average the time of query segments (25 frames).

Table 2. Results of video copy detection

	precision	recall	F1	Average computation
Proposed method	79.1%	88.3%	83.4%	39.3237s
Frame-based	60.5%	75.3%	67.1%	44.6943s

We can find from Table 2 that the proposed method has better result than typical frame-based method. As most of computation are finished off-line, and the features are much less than frame features, the process time is appropriate.

4 Conclusion

In this work, we propose an effective video copy detection system, which extracts spatio-temporal features instead of using spatial features and adding temporal information in a latter step. We abandon selecting frames (keyframes extracting by boundary-detection or uniformly sampling) but divide videos into segments with fixed length after preprocessing. Then a spatio-temporal interest point

detector is presented. Similar to the image case, we extract cuboid for each interest point containing most of the volume of data that contributed to the response function at that detected points. Then an extension of SIFT for cuboid representation is introduced. In order to further reduce the high-dimension of SIFT, we apply a PCA method. In the last step, we compare the similarity among segments using Euclidean distance, and determine the final detection result with a weighted-aggregate vote strategy. Experimental results show the effectiveness of our system. Moreover, we may fuse audio information to improve the detection.

Acknowledgement. This work is partly supported by NSFC (Grant No. 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Guidelines for the TRECVID 2011CD task Evaluation(OL) (2011), <http://www-nplpir.nist.gov/projects/tv2011/tv2011.html>
2. Kompatsiaris, Y., Merialdo, B., Lian, S.: TV Content Analysis: Techniques and Application. CRC Press, Taylor&Francis Group, Boca Raton, FL (2012)
3. Douze, M., Jegou, H., Schmid, C.: An Image-Based Approach to Video Copy Detection with Spatio-Temporal Post-Filtering. *IEEE Transactions on Multimedia* 12, 257–266 (2010)
4. Harvey, R.C., Hefeeda, M.: Spatio-Temporal Video Copy Detection. In: 20th ACM International Conference on Multimedia, pp. 35–46 (2012)
5. 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)
6. Law, J., Buisson, O., Gouet, V., Boujemaa, N.: Robust Voting Algorithm Based on Labels of Behavior for Video Copy Detection. In: 14th ACM International Conference on Multimedia, pp. 835–844 (2006)
7. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
8. Ke, Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 506–513 (2004)
9. Barrios, J.M., Bustos, B.: Competitive content-based video copy detection using global descriptors. *Multimedia Tools and Applications* 62, 75–110 (2013)
10. Laptev, I., Lindeberg, T.: Space-time interest points. In: 9th IEEE International Conference on Computer Vision, pp. 432–439. IEEE Press, New York (2003)
11. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. *IJCAI*, 674–679 (1981)
12. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–100 (2004)
13. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree Cover. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)

Improve Scene Classification by Using Feature and Kernel Combination

Lin Yuan, Fanglin Chen, Li Zhou, and Dewen Hu

College of Mechatronics and Automation, National University of Defense Technology,
Changsha, Hunan, P.R. China, 410073
dwhu@nudt.edu.cn

Abstract. Scene classification is an important issue in the computer vision field. In this paper, we propose an improved approach for scene classification. Compared with the previous work, the proposed approach has two processes to improve the performance of scene classification. First, feature combination is conducted to extract more effective information to describe characteristics of each category decreasing the influence of scale, rotation and illumination. Second, to extract more discriminative information for building a multi-category classifier, a kernel fusion method is proposed. Experimental results show that the use of the feature and kernel combination method can improve the classification accuracy effectively.

Keywords: Feature combination, Scene classification, Multi-resolution, Image categorization.

1 Introduction

Scene classification is a fundamental process in the computer vision system which enables the machine to analyze or understand the surrounding environment rapidly and effectively. So far, the automatic recognition and classification of scenes have been important issues in computer vision field, and they are widely used in computer vision applications, such as object recognition and detection [1] content-based image indexing and retrieval [2].

So far, there are lots of methods proposed to classify the natural scenes. The early scene classification models mainly concentrated on modeling scenes using low-level global statistical information [3], and these models are mainly utilized for binary classification problems [4], such as discrimination of indoor and outdoor scenes. These approaches have not been extended to the multi-category scene classification. Recently, the “bag of features” method has acquired huge success in the image analysis and classification [5]. But the conventional “bag of features” approach uses only one local descriptor to represent local regions and only uses the single resolution feature channel. The simplex descriptor is not powerful enough to describe the local regions and the combined feature can contain more effective information to represent the images. Because the multi-category is a non-linear classification problem, the kernel method is widely used

in the scene classification. But the single kernel has limited power to distinguish and the kernel fusion can make the classifier have stronger discriminating power for multi-category [6].

In this paper, we use the methods of feature combination in feature processing and kernel fusion in kernel calculating. Then we use the two method to improve the conventional the “bag of features” approach to classify the scene images. We first create three resolutions for each image, calculate the features such as WHGO (weighted histogram gradient orientation) [7], LBP (local binary pattern) [8], SIFT (scale invariant feature transform) [9] which have obtained better performances for scene classification and concatenate the different features as a descriptor in each resolution. Then we use the k -means cluster algorithm to calculate the code book, forming the visual dictionary. After that we use the dictionary to represent the images in the three resolutions, and calculate kernels respectively. In the kernel combination step, the weights of different kernels are learned adaptively and the optimal parameters are obtained to get the best performance.

2 Approach of This Paper

2.1 Feature Combination

Our approach works by creating multiple resolution images and partitioning them into sub-regions at different scales. We create three resolution images through sub-sampling. We represent each sub-region with WHGO, LBP and SIFT descriptors, and concatenate each descriptor of all local regions in the same resolution image to form an image representation in this resolution. We regard each descriptor of each resolution image as a feature descriptor. The final feature is combined as:

$$d(i, j) = \{d_1(i, j), d_2(i, j), d_3(i, j)\}, \quad (1)$$

where $d(i, j)$ means the final descriptor of the local regions with the center (i, j) and $d_k (k = 1, 2, 3)$ means the three descriptors correspond to the three different resolution images.

2.2 Kernel Combination

After executing the cluster algorithm and representing by the code book, we can obtain three resolution feature channels. We regard each resolution image as a feature channel in our approach, and we use the feature combination method mentioned above to combine the features of different resolutions. The multiple category scenes are classified with an SVM (support vector machine) with the training samples to build a classifier using the one-versus rest rule and all parameters of SVM classifiers are obtained through parallel training. The conventional classification uses only one kernel to map the low dimensional feature to high

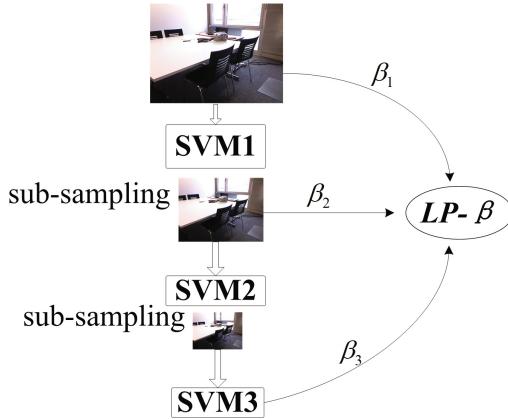


Fig. 1. Combination of multi-resolution feature channels

dimensional one, but the power to distinguish is limited. In this paper, we improve the traditional kernel calculating method by fusing the different resolution kernels. Shown as the Fig. 1, after all SVM classifiers of all resolution feature channels are obtained by training, the different resolution feature channels are combined by using the $LP\text{-}\beta$ approach introduced by Gehler [10]. The decision function for a test image x has the following form:

$$y(x) = \arg \max_{c=1,2,\dots,C} \sum_{ch=1}^F \beta_{ch} \left(K_{ch}(x)^T \alpha_{ch,c} + b_{ch,c} \right), \quad (2)$$

where $K_{ch}(x) = (K_{ch}(V_1, V_x), K_{ch}(V_2, V_x), \dots, K_{ch}(V_N, V_x))^T$ means the kernels of different channels, the V_i means the feature vector of the i class in the training images and the V_x means the feature vector of the test images. N is the number of the classifications, F is the number of channel resolutions, α are the weight parameters got by training, and b is the learned threshold parameters of all resolution feature channels. x is the testing image, and y is the corresponding class label of x . β is the mixing coefficients and can be learned by the following form:

$$\min_{\beta, \xi, \rho} -\rho + \frac{1}{vN} \sum_{i=1}^N \xi_i, \quad (3)$$

$$s.t. \sum_{ch=1}^F \beta_{ch} f_{ch,y_i}(x_i) - \arg \max_{y_j \neq y_i} \sum_{ch=1}^F \beta_{ch} f_{ch,y_j}(x_i) + \xi_i \geq \rho, i = 1, 2, \dots, N, \quad (4)$$

$$\sum_{ch=1}^F \beta_{ch} = 1, \beta_{ch} \geq 0, ch = 1, 2, \dots, F, \quad (5)$$

where $\{(x_i, y_i)\}_{i=1,2,\dots,N}$ is the training set, x_i is the i th training image and $y_i \in \{1, 2, \dots, C\}$ is the corresponding class label of x_i .

2.3 Bag of Features

The proposed approach used feature and kernel combination to improve the conventional “bag of features” method. Fig. 2 is the flow chart of our approach. In this paper, the proposed approach improves the traditional “bag of features” method. In the training step, it firstly creates three resolution images and extracts local feature descriptors in different scales for each resolution. After extracting the features, it builds the code book by using the k -means cluster algorithm. Then it uses the code book to represent the images in each resolution. Finally it combines the three resolution feature channels. In the testing process, it firstly extracts the features of images as the training process, and secondly it uses the code book obtained in the training process to represent the testing image. Finally the testing image is assigned the label of the SVM classifier.

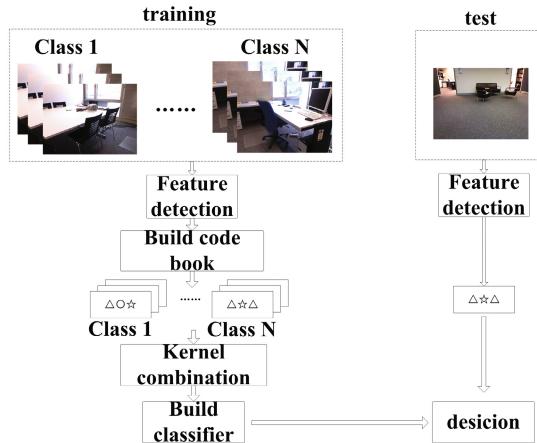


Fig. 2. Flow chart of our approach

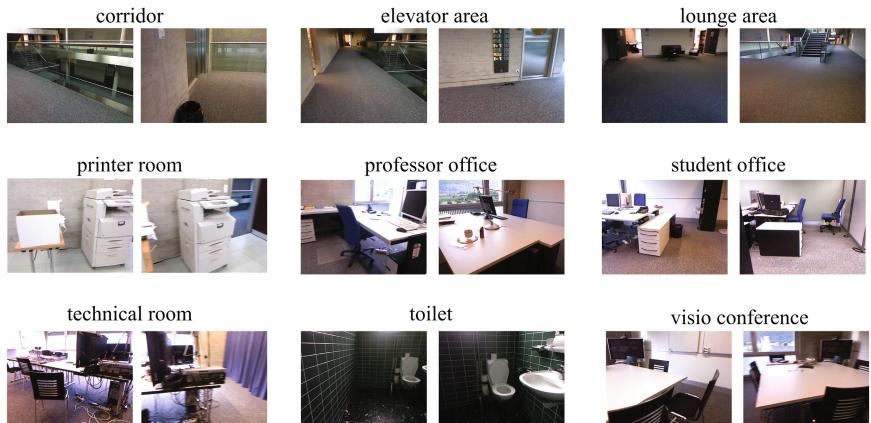
3 Data Sets and Experimental Approach

3.1 Data Sets

The scene classification approach of this paper is tested in the ImageCLEF2012 (Robot Vision Challenge) data set. The ImageCLEF is organized by Royal Institute of Technology, IDIAP Research Institute of Switzerland and Georgia Institute of Technology. The task of this competition is to localize the robot and the key point is to finish visual place classification.

Table 1. The component of the ImageCLEF2012

	corridors	elevator	lounge	printer	professor	student	technical	toilets	visio
	areas	areas	rooms	offices	offices	rooms	rooms		conferences
training	882	224	797	184	655	1052	402	198	186
testing	498	152	452	80	336	599	96	240	79

**Fig. 3.** Example images of the ImageCLEF2012 data sets

In ImageCLEF2012, the data set contains 9 categories, including training sets and testing sets. The Table. 1 gives the details of the image data set and the Fig. 3 shows example images from the ImageCLEF2012 data sets.

3.2 Experiment Approach

The task of scene classification is to assign a class label of a number of categories to each testing images. In the experiments, we use the regulation of the competition to test our approach. In the competition, if the assigned class label equals to the actual label, then the final score plus 1, otherwise minus 1. The number of test images is 2532, so the highest score should be 2532. The score can reflect the classification accuracy: the higher score, the higher accuracy.

To test the approach of feature combination, we use WHGO, LBP and SIFT to describe the local regions and use the mixing coefficient β to combine the feature channels respectively and then concatenate the three descriptors as a final descriptor with the same approach.

4 Results

4.1 Relationship between Mixing Coefficients and Final Score

The classification accuracy is related with the mixing coefficient β because the three different resolution feature channels are combined to get the classifier in this paper. The mixing coefficients β in this paper are learned by the $LP\text{-}\beta$ approach introduced above. To describe the influence of each resolution feature channel on the final score more clearly, we use 21 values for β_1 , β_2 from 0 to 1 in an interval of 0.05, and $\beta_1 + \beta_2 + \beta_3 = 1$ is ensured. The Table. 2 gives the mixing coefficients with the highest score and the Fig. 4 shows the trend with the different mixing coefficient β .

Table 2. The mixing coefficients for the best performance

Features	Highest score	β_1	β_1	β_1
WHGO	1474	0.7	0.05	0.25
LBP	1382	0.3	0.15	0.55
SIFT	1474	0.4	0.3	0.3
Feature combination	1714	0.65	0.1	0.25

4.2 The Results after Feature Combination

We use the feature combination approach introduced above in our experiments and the results demonstrate that the feature combination is surely useful for improving the performance. The multi-category scenes classifiers employed in this work are based on a nonlinear SVM with a χ^2 kernel. The Table. 3 gives the contrast of the score between feature combination and non-combination approach.

From the Table. 3 we can easily find that the score increases about 150 only by using feature combination, about 216 scores only by using kernel combination and about 240 scores by using the mixed method. From the result we can obtain the conclusion that the feature combination of different resolution feature channels is useful for improving the performance of the scene classification. This means our approach is effective.

Table 3. The contrast of the score between feature combination and non-combination approach

Score	WHGO	LBP	SIFT	feature combination
Kernel combination	1474	1382	1474	1714
Single resolution	1258	1076	1254	1408

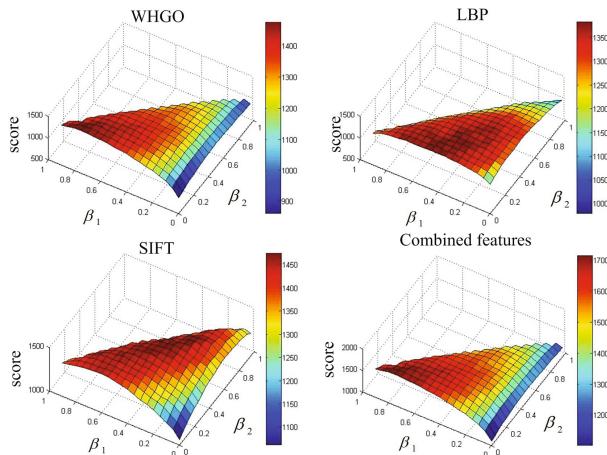


Fig. 4. Relationship between mixing coefficients β and final score of each descriptor

5 Conclusions

This paper has presented a novel method by using feature combination approach on “bag of features” approaches for scene categories based. Our method works by partitioning images of multi-resolution and describing the local regions in each resolution image to build a feature channel. The method has achieved promising results on the ImageCLEF2012 data set through the combination of all resolution feature channels. Our experiments show that the feature combination of different resolution feature channels is meaningful for improving the performance of the scene categories. This means our approaches are effective.

Acknowledgments. This work was supported by the National Natural Science Foundation of China(Grant Nos. 61203263), the National Basic Research Program of China(No. 2013CB329401, 2011CB707802), New Century Excellent Talents in University(NCET-08-0147), Hunan Provincial Innovation Team Project.

References

1. Torralba, A.: Contextual priming for object detection. *International Journal of Computer Vision* 53(2), 169–191 (2003)
2. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision* 72(2), 133–157 (2007)
3. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
4. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: *Proceedings of IEEE International Workshop in Content-based Access of Image and Video Database*, pp. 42–51 (1998)

5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
6. Zhao, C., Liu, C., Lai, Z.: Multi-scale gist feature manifold for building recognition. *Neurocomputing* 74(17), 2929–2940 (2011)
7. Li, Z., Dewen, H., Zongtan, Z., Zhaowen, Z.: Natural Scene recognition using weighted histograms of gradient orientation descriptor. *Front. Electr. Electron. Eng. China* 6(2), 318–327 (2011)
8. 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)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
10. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: Proceedings of IEEE 12th International Conference on Computer Vision, pp. 221–228 (2009)

Horror Text Recognition Based on Generalized Expectation Criteria

Guoqi Liu¹, Bing Li², Weiming Hu², and Jinfeng Yang¹

¹ Civil Aviation University of China

² Institute of Automation, Chinese Academy of Sciences, Beijing, China
liuguoqi0248@126.com, jfyang@cauc.edu.cn,
{bli,wmhu}@nlpr.ia.ac.cn

Abstract. Along with the growth of Internet, Web is glutted with more and more illegal and harmful information, such as pornography, violence, horror information. For a long time, researchers pay little attention to horror information relative to pornography. Horror information as well as pornography harms youngsters' health seriously. In order to recognize horror information, in this paper, we propose a horror text recognition algorithm based on two classifiers, in which one is text title classifier, and the other is text content classifier based on generalized expectation (GE) criteria. A generalized expectation criterion is a term in a parameter estimation objective function that assigns scores which can express preferences to values of a model expectation. In this paper, this parameter estimation objective function is used for measuring the correlation between features and sentiment labels.

Keywords: horror text, recognition, generalized expectation criteria.

1 Introduction

With the increasing use of network by the general public, a great quantity of horror information which is easily accessed by youngsters has exploded into yet unseen numbers. Many psychological and physiological researches indicate that too much horror information can seriously affect youngsters' health. Rachmans research [1] shows that horror information is one of the most important factors for phobias. Field et al. [2] further point out that horror information can increase behavioral avoidance as well as fear beliefs. The experiments of King et al. [3] further indicate that 88% youngsters ascribe their phobias to horror information acquisition.

With the explosion of horror information on the Web, recognizing and filtering the horror information from the Web is becoming an increasingly urgent task. We have observed that horror information can be divided, according to form of existence, into three categories: horror text, horror image and horror video. In this paper, we focus on horror text recognition.

1.1 Related Work

Horror texts are to elicit the emotions of fear, horror and terror from readers. Horror text recognition can be viewed as a sentiment analysis task. The classical work in

sentiment analysis [4, 5] view sentiment classification as a text classification problem where an annotated corpus with documents labeled with their sentiment orientation is required to train the classifiers. Theresa et al. [6] propose an approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions, achieving results that are significantly performed well. In paper [7], a method based on K-nearest neighbor for sentiment analysis is proposed. It primarily confirms the sentiment of each paragraph, and then confirms the whole text's sentiment. Wang et al. [8] provide a method of text sentiment classification based on weighted rough membership. In the method, a text expression model is established based on two-tuples attribute (feature, feature orientation intensity), by introducing feature orientation intensity into vector space representation.

1.2 Our Work

Sentiment classification aims to automatically predict sentiment polarity and usually we concentrate on discriminating between positive and negative or thumbs up and thumbs down. Many methods focus on sentiment analysis about news-review domain and product-review domain. However, to the best of our knowledge, there is nearly little research on the horror text recognition. The differences between horror text recognition and general sentiment classification are summarized in table 1. Through the former experiments we have done, we draw a conclusion that horror text possess two characteristics. On one hand, some horror texts obviously contain feature words, for instance, “bloody”, “ghost”, “skeleton”, and so on. On the other hand, there are some texts that contain no explicit words that are obviously horrible. Therefore, the traditional methods based on topic sentences are unsuitable for horror text recognition. Besides, the titles of horror texts are helpful items and are incorporated into horror text classification. We expect that it can eliminate the complex analysis for the content of texts and improve the accuracy of sentiment analysis.

Table 1. Distinctions between horror text classification and general sentiment classification

Contrast items	General sentiment classification	Horror text recognition
State of research	A lot	A little, be ignored usually
References	Many	Few
Features extraction	Easy	Difficulty
Application domain	News-review and product-review	Horror text recognition
Distinction degree of the title	Inconspicuous	Obvious
Classificatory level	Low-level	High-level

The rest of the paper is structured as follows. The proposed method is introduced in Section 2. The experimental results are presented in Section 3. Finally, Section 4 concludes the paper.

2 Proposed Method

In this paper, we take advantage of both the title and the content of a text, and propose a novel framework for recognizing horror texts as it presents in figure 1. In our framework, horror texts are divided into two categories, each of which is handled by the corresponding classifier. The recognition result is determined by the fusion of the results obtained by the title classifier and the content classifier.

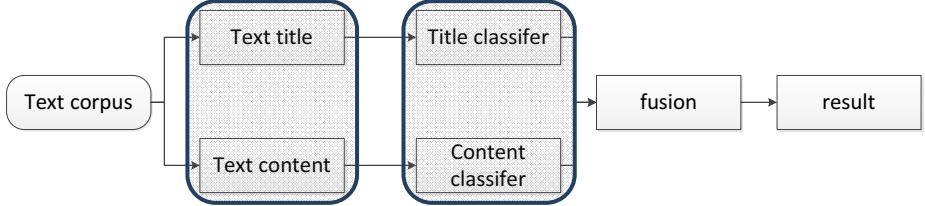


Fig. 1. Overview of our work

2.1 Content Classifier

In this section, we describe generalized expectation criteria and derive the parameter estimation objective function we use to extract content features. Those extracted content features are used to train SVM which is employed to classify text content.

A generalized expectation (GE) criterion [9] is a term in a parameter estimation objective function that assigns scores which can express preferences to values of a model expectation. Given a score function G , an empirical distribution \tilde{P} , a function f , and a conditional model distribution P parameterized by θ as well as $E_\theta[f(X, Y)]$ the expectation of function f , the value of a generalized expectation (GE) criterion is:

$$G(E_\theta[f(X, Y)]) = G(E_{\tilde{P}}[E_{P_\theta(Y|X)}[f(X, Y)]])$$

One specific type of score function G is some measure of distance between the model expectation and a reference expectation. Given some distance function $\Delta(\cdot, \cdot)$, a reference expectation \hat{f} , this criterion is:

$$G(E_\theta[f(X, Y)]) = \Delta(\hat{f}, E_{\tilde{P}}[E_{P_\theta(Y|X)}[f(X, Y)]])$$

In this paper, we use the Kullback–Leibler divergence (KL) for $\Delta(\cdot, \cdot)$. Therefore, function G can be defined:

$$G(E_\theta[f(X, Y)]) = KL(\hat{f}, E_{\tilde{P}}[E_{P_\theta(Y|X)}[f(X, Y)]])$$

Sentiment labels set S is denoted by $S = \{\text{horrible}, \text{non-horrible}\}$. Texts collection D is denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n\}$ where the bold-font variables denote the vectors. Each text in D is a sequence of words denoted by vector $\mathbf{w}_i = \{t_1, t_2, \dots, t_m\}$.

We define two-tuples attribute (word, class label) feature function as $f(X, Y)$ and let it be $f(X, Y) = f_{jk}(\mathbf{w}, s) = \sum_{i=1}^D \delta(s_i = j)\delta(k \in \mathbf{w}_i)$, where $\delta(x)$ is an indicator

function which takes a value of 1 if x is true, 0 otherwise, and s_i is some sentiment label, k is some feature. And then, the expectation of the features can be expressed as:

$$E_\theta[f(\mathbf{w}, s)] = E_{\widetilde{P(\mathbf{w})}} \left[E_{P_\theta(s|\mathbf{w})} [f(\mathbf{w}, s)] \right] = \widetilde{P(\mathbf{w})} \cdot P_\theta(s|\mathbf{w}) \cdot f(\mathbf{w}, s)|_{\mathbf{w}=\mathbf{w}_i, s_i=j, k \in \mathbf{w}_i}$$

The probability of s conditioned on \mathbf{w} is given by $P_\theta(s|\mathbf{w}) = \frac{\exp(\sum_i \theta_{si} t_i)}{Z(\mathbf{w})}$ where $Z(\mathbf{w}) = \sum_s \exp(\sum_i \theta_{si} t_i)$ is a normalizer that makes sure $\sum_s P_\theta(s|\mathbf{w}) = 1$ and $\widetilde{P(\mathbf{w})} = \widetilde{P(t_1)} \cdot \widetilde{P(t_2)} \cdots \widetilde{P(t_n)}$ is the empirical distribution of \mathbf{w} in texts collection D . Model parameters θ_{si} can be estimated by using improved iterative scaling (IIS). So:

$$E_\theta[f(\mathbf{w}, s)] = \widetilde{P(t_1)} \cdot \widetilde{P(t_2)} \cdots \widetilde{P(t_n)} \cdot \frac{\exp(\sum_i \theta_{si} t_i)}{\sum_s \exp(\sum_i \theta_{si} t_i)} \cdot \sum_{i=1}^D \delta(s_i = j) \delta(k \in \mathbf{w}_i)$$

Plug equation $E_\theta[f(\mathbf{w}, s)]$ into $G(E_\theta[f(\mathbf{w}, s)])$, we get:

$$G(E_\theta[f(\mathbf{w}, s)]) = \sum_s \widehat{f}_{sk} \cdot \log \frac{\widehat{f}_{sk}}{E_\theta[f_{sk}(\mathbf{w}, s)]}$$

We select k as a feature when $G(E_\theta[f(\mathbf{w}, s)])$ is less than a threshold. We use the method of improved iterative scaling (IIS) to estimate parameters.

2.2 Title Classifier

Let $C = \{c_1, c_2 \dots c_l\}$ be the set of title categories, where l is the number of title categories (in this paper, $C = S$, $l=2$). The title of a text can be expressed as $(a_1, a_2 \dots a_n)$, where n is the number of components in the vector and a_i is the i th word in the title. The probability $P(c_j|a_1, a_2 \dots a_n)$ that the vector belongs to category c_j is determined by the following equation:

$$P(c_j|a_1, a_2 \dots a_n) = \frac{P(a_1, a_2 \dots a_n|c_j)P(c_j)}{\sum_r^l P(a_1, a_2 \dots a_n|c_r)P(c_r)}.$$

The term $P(c_j)$ is estimated by the frequency of the samples belonging to c_j in the training data. The estimation $P(a_1, a_2 \dots a_n|c_j)$ is difficult because the data are sparsely distributed in a high-dimensional space unless there are a great many of testing data.

We investigate the titles and statistics show that words in titles are relatively independent. And so $P(c_j|a_1, a_2 \dots a_n) = \frac{P(c_j) \prod_i P(a_i|c_j)}{\sum_r^l P(c_r) \prod_i P(a_i|c_r)}$. The term $P(a_i|c_j)$ is easily estimated by accounting the feature word a_i in category c_j .

Given a title t , we check whether the probability $P(c_j|t)$ that the input title t belongs to the horror title category exceeds a predefined threshold. If so, the title t is classified as horror title; otherwise, the title is classified as non-horror title.

2.3 Fusion

In this paper, we consider a text is horrible when both of the content and title of the text are classified as horrible, and a text is non-horrible when both of the content and title of the text are classified as non-horrible. Besides, if $T \geq 1$, the text is classified as horrible, and if $1 \geq T \geq 0$, the text is classified as non-horrible. T is a decision factor.

We define four statistical features of the classifier: P_1 , P_2 , P_3 and P_4 . The probability P_1 means that the content of a horror text is mistakenly classified as non-horror content and the probability P_2 means that the title of a horror text is mistakenly classified as non-horror title. The probability P_3 means that the content of a non-horror text is mistakenly classified as horror content and the probability P_4 means that the title of a non-horror text is mistakenly classified as horror title. The probability P_1 as well as P_2 can be estimated statistically by counting the number of non-horror contents mistakenly classified by the classifier in a set of horror text. The probability P_3 as well as P_4 can be estimated statistically by counting the number of horror contents mistakenly classified by the classifier in a set of non-horror text. In this paper, we set $P_1=0.15$, $P_2=0.137$, $P_3=0.2$ and $P_4=0.2$ according to the statistic on training set.

Let Q represent the event that the text is a horror text and $\neg Q$ represent the event that the text is a non-horror text. Let r represent the event that one of the title and content of a text is classified as horror; the other is classified as non-horror. Then, the equations below are obtained, 1) $P(r|Q) = (1 - P_1)P_2 + P_1(1 - P_2)$, 2) $P(r|\neg Q) = (1 - P_3)P_4 + P_3(1 - P_4)$

We introduce a decision factor T , which is the ratio of the two posteriori probabilities, $P(r|Q)$ and $P(r|\neg Q)$:

$$T = \frac{P(Q|r)}{P(\neg Q|r)} = \frac{P(r|Q) \cdot P(Q)}{P(r|\neg Q) \cdot P(\neg Q)} = \frac{(1 - P_1)P_2 + P_1(1 - P_2)}{(1 - P_3)P_4 + P_3(1 - P_4)} \cdot \frac{P(Q)}{P(\neg Q)}$$

If $T \geq 1$, the text is classified as horrible, and if $1 \geq T \geq 0$, the text is classified as non-horrible.

The remaining problem is to confirm the a priori probabilities $P(Q)$ and $P(\neg Q)$. We search “fiction” in *baidu* search engine and obtain 24,160,000 webpages, and then we search “horror fiction” in *baidu* search engine and obtain 17,660,000 webpages. Accordingly, the probability $P(Q)$ is set equal to $P(Q) = 17660000/24160000 = 0.73$, so $P(\neg Q) = 1 - P(Q) = 0.27$. One point should be mentioned here: there are many kinds of fiction, and obviously that $P(Q) = 0.73$ is a very high value. We may be confused. But in this paper, all of the texts are from Web and we aim to structure an effective horror text filtering tool for Web horror information. Therefore, the method of estimating $P(Q)$ is feasible.

3 Experiments

We collected a large number of corpuses that consists of 1000 horror texts and 1000 non-horror texts. The horror texts including thriller and ghost are downloaded from some horror Webs and horror fiction forums, and the non-horror texts contain science fictions, military fictions as well as the field of news, education, and life. We divide the horror texts and non-horror texts into five parts respectively on average, and perform

5-fold cross validation. Four out of five horror texts and non-horror texts are used for training data, and the rest are used for testing date. We perform using different feature vector dimensions: 300, 600, and 800 shown in figure 2. And then, we compare the feature extraction method of lexicon labeling (LL), information gain (IG) with GE we propose in this paper in different feature vector dimensions, and the results are shown in figure 3, figure 4, and figure 5.

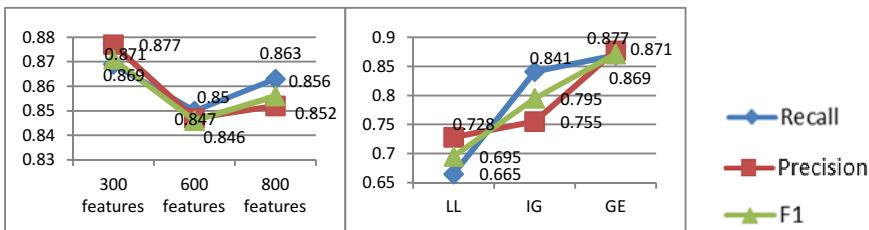


Fig. 2. Performance of different feature vector dimensions for GE

Fig. 3. Comparison in feature extraction methods for 300 feature vector dimensions

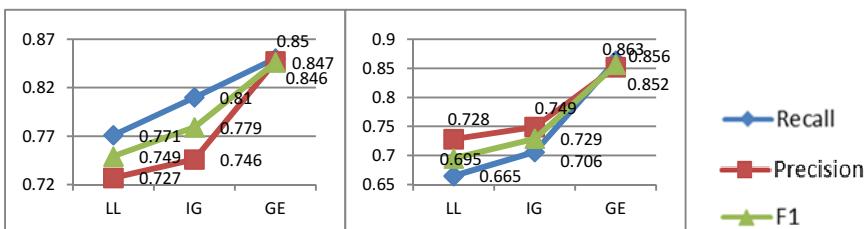


Fig. 4. Comparison in feature extraction methods for 600 feature vector dimensions

Fig. 5. Comparison in feature extraction methods for 800 feature vector dimensions

From figure 2, compared with the other three conditions, the 300 feature vector dimensions achieve better performance and the recall, precision, F₁ are 86.9%, 87.7% and 87.1% respectively. We have observed that the curves are not monotone increasing along with the increasing of feature dimensions. One reason is possible that the curves may be inflected by semantic analysis and threshold setup of feature extracting. However, both of the factors are not involved in this paper, and will be discussed in the future work. Another possible reason is the corpuses we collect here are short texts. Those texts mapped to high dimension feature vector space are sparse.

From figure 3 to figure 5, we compare lexicon labeling, information gain with GE we propose in different feature vector dimensions and the method we propose in this paper is effective and comparable with the other two methods. In lexicon labeling method, we asked three M.E. to choose good indicator words for horror text content and this method result in the worse performance.

4 Conclusion

In this paper, we have proposed an effective approach to solve the problem of horror text recognition. Generalized expectation (GE) criterion is introduced into the horror text recognition and we use GE to extract text content features. Then, we employ support vector machine (SVM) to classify text content. A native Bayesian classifier is also used for title classification. At last, the recognition result is determined by the fusion of the results obtained by the title classifier and the content classifier.

Acknowledgments. This work is partly supported by NSFC (Grant No. 60935002, 61005030), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012503, 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Rachman, S.: The conditioning theory of fear acquisition: A critical examination. *Behavior Research and Therapy* 15(5), 375–387 (1977)
2. Field, A.P., Lawson, J.: Fear information and the development of fears during childhood: Effects on implicit fear responses and behavioral avoidance. *Behavior Research and Therapy* 41(11), 1277–1293 (2003)
3. King, N.J., Eleonor, G., Ollendick, T.H.: Etiology of childhood phobias: Current status of rachmans three pathways theory. *Behavior Research and Therapy* 36(3), 297–309 (1998)
4. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics, pp. 440–447 (2007)
5. Narayanan, R., Liu, B., Choudhary, A.: Sentiment analysis of conditional sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 180–189 (2009)
6. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: EMNLP, pp. 347–354 (2005)
7. Fan, N., An, Y.S., Li, H.X.: Research on analyzing sentiment of texts based on k-nearest neighbor algorithm. *Computer Engineering and Design* 33(3) (2012)
8. Wang, S.G., Li, D.Y., Wei, Y.J.: A method of text sentiment classification based on weighted rough membership. *Journal of Computer Research and Development* 48(5), 855–861 (2011)
9. McCallum, A., Mann, G., Druck, G.: Generalized expectation criteria. Technical Report 2007-62, University of Massachusetts, Amherst (2007)

A Representative-Sequence Based Near-Duplicate Video Detection Method

Chen Shi, Li Zhuo, Yingdi Zhao, and Yuanfan Peng

Beijing Signal and Information Processing Lab,
Beijing University of Technology, Beijing, China
shichen@emails.bjut.edu.cn, zhuoli@bjut.edu.cn,
kate333333@emails.bjut.edu.cn, pengyuanfan@sina.com

Abstract. This paper presents a method of near-duplicate video detection based on representative-sequence. Firstly, the video is divided into different scenes according to the variance of the Chi-squared color histogram and Grayscale OM (Ordering Measure), and then the frames of a single scene are preprocessed by decreasing the frame rate to a particular rate and detecting the occurrence of black-sides. For each scene, every certain number of frames are merged into one TIRI frame, and then weighted grayscale mutual information is calculated to select the representative frames from the TIRI frames. Next, several features are extracted from the representative frames for matching, including color, edge and grayscale features. Finally, whether or not the video is a near-duplicate is determined by the match degree of the characteristics. Experimental results show that the algorithm presented by this paper possesses good robustness and detection capabilities.

Keywords: Near-duplicate video detection, TIRI, Ordering measure, MPEG-7.

1 Introduction

With the development of science and technology, video has become an important information carrier in daily life. Multimedia processing technology greatly facilitate the users to copy and edit a video clip at ease, which at the same time brings the problem of copyright protection of digital media. Thus, video copy detection technology[1] has been carried out. The video copy detection technology is a method to determine whether the two videos with different manifestations are of the same content by matching the features of the video contents[2].

From the perspective of data processing, existing video copy detection techniques can be divided into two types: compressed-domain methods and pixel-domain methods.

The compressed-domain methods directly extract information from the compressed video bitstream without decoding, and therefore generally have a faster detection speed[3]. However, the types of feature available are limited, and the methods are usually restricted to certain video formats.

The pixel-domain methods extract features from pixel values, which generally can be classified into three kinds, namely the spatial features, the temporal features and the spatial-temporal features.

Spatial features possess the most types of features. They are extracted from a single video frame, and can be further divided into global and local features. Global features are extracted by the processing and induction of the global statistics of low-level features. Common global features include the block-based ordering measure(OM)[4,5], color histogram^[3] and features based on color shift and the centroid[6]. Global features can be extracted at a high speed, and also meet the needs of accuracy. Local features are extracted in the local area of feature points detected after image partitioning. Common local features include SIFT[7] and CS-LBP[8]. Though local features improve the detection accuracy to some extent, the computation complexity brought by them is often too high to meet the real-time processing requirements.

In general, spatial features describe the spatial relationship of the features, while ignoring the temporal information contained in the video. As a result, researchers also paid attention to temporal features. Common temporal features include the moving trajectory[9], the motion vector[10] and temporal-spatial slices[11]. These features characterize the temporal information of the video, yet provide poor results when applied to videos of a short duration.

Compared with them, the spatial-temporal features take the advantage of both spatial features and temporal features. Mani proposed a method of generating temporally informative representative images (TIRI)[12], which contains the spatial-temporal information of a short section of the video. However, as the method did not take scene change into account, the TIRI could not effectively represent the video information. Moreover, the overlapping method generates a great number of TIRIs, leading to a large quantity of redundant information.

To solve these problems, a representative-sequence based near-duplicate video detection method is proposed in this paper. First, the method to generate TIRI frames is improved by scene segmentation and preprocess. After that, features based on MPEG-7 standard are extracted. At last, a three-level similarity matching mechanism is proposed to realize the near-duplicate video detection.

The rest of this paper is organized as follows: Section II briefly introduces the framework of the proposed near-duplicate video detection scheme; Experimental results are presented and analyzed in Section III; Finally conclusion is drawn in Section IV.

2 A Near-Duplicate Video Detection Mechanism Based on Representative-Sequence

The proposed near-duplicate video detection method based on representative-sequence includes three parts, namely the generation of representative-sequence, the extraction of features and the similarity matching part.

2.1 Representative-Sequence Generation Algorithm

The diagram of the generation of representative-sequence is shown in Fig.1. The video sequences are first segmented into scenes. Then, the frames of a single scene are under a series of preprocess. After that, every certain number of frames in the same scene are merged into one TIRI frame. At last, weighted grayscale mutual information is calculated to select the representative frames (RF) from the TIRI frames.

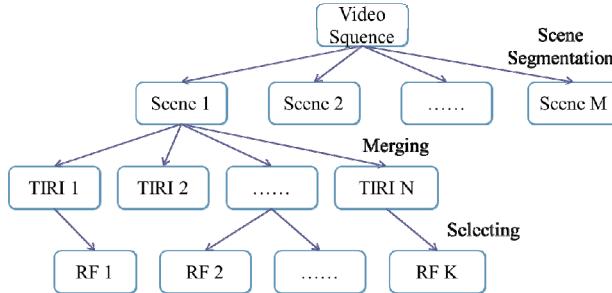


Fig. 1. A diagram of the generation of the representative- sequence

Scene Segmentation. The process of scene segmentation is shown in Fig.2. Video sequences are divided into different scenes according to the variance of the Chi-squared color histogram and grayscale OM sequences.

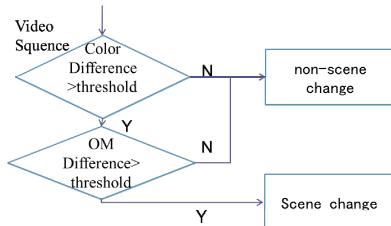


Fig. 2. The scene segmentation flowchart

10	40	20
50	70	90
30	80	60

$$\begin{bmatrix} 1 & 4 & 2 \\ 5 & 7 & 9 \\ 3 & 8 & 6 \end{bmatrix}$$

Fig. 3. The diagram of the generation of OM

First, the first frame is set as the reference frame. Then, the variance of the Chi-squared color histogram between the following frames and the reference frame are calculated as (1):

$$Diff(F_i, F_j) = \sum_{k=0}^M \frac{(H(F_i, k) - H(F_j, k))^2}{H(F_i, k) + H(F_j, k)} \quad (1)$$

where F_i, F_j refers to the two frames, H refers to the color histogram. If the result is bigger than the threshold, the difference between the grayscale OM sequences will be calculated.

A grayscale OM sequence is defined as Fig.3. Divide the frame into 9 blocks and calculate the average grayscale of each block. The sorting order from the smallest to the biggest forms the OM sequence.

If the Hamming Distance between the current frame and the reference frame is larger than the threshold, a new scene is detected. Then, the current frame is set as the new reference frame. The same process continues until the end of the video sequence.

Generation of TIRI Frames. After scene segmentation, the frames of a single scene are preprocessed by decreasing the frame rate to a particular rate and detecting the occurrence of black-sides. For each scene, every certain number of frames are merged into one TIRI frame, as shown below:

$$o_{m,n} = \sum_{k=1}^L \omega_k l_{m,n,k} \quad (2)$$

where L is the number of the merged frames, $l_{m,n,k}$ refers to the pixel value at position (m,n) in the k^{th} frame. In this paper, ω_k is set as 0.2.

Selection of Representative Frames Based on Mutual Information. After the generation of TIRI frames, a series of frames are obtained which contain redundant information. Thus, a method of selection of representative frames based on mutual information is proposed.

Given that important information is usually expressed in the middle of the picture, a block-based mutual information calculating method is proposed here.

The original mutual information is defined as:

$$MI_k(A, B) = \sum_{a,b} p_{AB}(a, b) \log \frac{p_{AB}(a, b)}{p_A(a) \cdot p_B(b)} \quad (3)$$

where $p_A(a), p_B(b)$ refers to the value of the normalized grayscale histogram at scale a, b in frame A, B respectively. $p_{AB}(a, b)$ refers to the value of the joint histogram. Applied to a block-based method, the mutual information is defined as:

$$WMI = \sum_k \omega_k MI_k(A, B) \quad (4)$$

where the values of ω_k is shown as Fig.4.

	0.125	
0.125	0.5	0.125
	0.125	
	0.125	

Fig. 4. A diagram of the weights of different parts

The selection process is as follows: select the first TIRI frame as a representative-frame, and set it as the reference frame. Then calculate the WMI value between the following TIRI frames and the reference frame. Once the value is larger than the threshold, the current TIRI frame will be selected as a new representative frame and will be set as the new reference frame. The process continues until the end of the scene.

2.2 Feature Extraction Algorithm

To avoid excessive computational complexity, this paper extracts color, edge and grayscale global features from the representative frames.

Color Feature Extraction. The color layout descriptor defined in MPEG-7[13] is selected as the color feature. First, the frame is divided into 64 blocks and a single representative color, which is the average of the pixel colors, is selected from each of the blocks. Then, transform the tiny image of 8×8 by 8×8 DCT so that 3 sets of 64 DCT coefficients are obtained. After zigzag-scanning and nonlinear quantization, 12 coefficients including 6 for luminance and 3 for each chrominance are selected as the color feature.

Edge Feature Extraction. The edge feature is with reference to the edge histogram descriptor defined in MPEG-7. Edge histogram is extracted by dividing the picture into 16 blocks and counting the percentage of different types of edge in a single block. The types of edge are defined as horizontal, vertical, 45° diagonal, 135° diagonal and non-directional.

The edge feature proposed in this paper is extracted as follows. First, divide the frame into 16 blocks. Then, count the percentage of different types of edge in a single block. After that, sort the percentages from the smallest to the biggest. At last, the median is used to change the sorted sequence to a binary sequence, so that the OM sequence as the edge feature can be obtained.

Grayscale Feature Extraction. Four grayscale features are employed in this paper. First, an OM sequence is calculated. Each frame is divided into 9 blocks, and the average value of each block is calculated. Then, compare the value of each block with the value of the middle block clockwise. If the former one is bigger, set it as “1”, else set it as “0”.

Then, the mean, standard deviation and contrast of gray-scale are calculated as the last three grayscale features.

2.3 Similarity Matching Algorithm

In order to realize the detection of near-duplicate videos, the matching algorithm in this paper is divided into three levels, including the frame level, the scene level and the video level. The total similarity between two video sequences is determined through the results of these levels.

Frame-Level Similarity Matching. The similarity between two representative frames is determined by the matching results of all the features extracted from the frames. Features of color, edge and grayscale are exploited from the frames in this paper. Depending on its form, the similarity of each kind of feature is determined in a unique way. If the distance between features of the same kind is beyond the empirical threshold, then that feature will be judged as “not similar”.

As for the color feature, the distance between two color layout descriptors is defined in MPEG-7 as:

$$D = \sqrt{\sum_i \omega_{yi} (DY_i - DY_i')^2} + \sqrt{\sum_i \omega_{bi} (DCb_i - DCb_i')^2} + \sqrt{\sum_i \omega_{ri} (DCr_i - DCr_i')^2} \quad (5)$$

where DY_i , DCb_i and DCr_i refers to the i^{th} DCT coefficient of Y, Cb and Cr component, respectively. Their weights, ω_{yi} , ω_{bi} and ω_{ri} , are defined in MPEG-7.

As for the edge feature and the grayscale OM sequence, Hamming Distance is employed for similarity matching. The absolute value of difference is selected for the mean of grayscale. At last, the similarity of the standard deviation and the contrast of grayscale are determined by the ratio of the corresponding absolute differences to the feature with a smaller value.

After the matching of features, the frame-level matching is realized by a voting mechanism, which is shown in Fig.5.

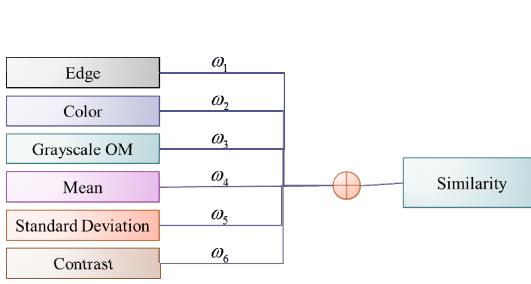


Fig.5. A sketch map for a weighted image matching

	V_t	T_1	T_2	T_3	T_4	T_5	T_6
Q_q	0	0	0	0	0	0	0
Q_1	0	0	0	0	1	1	1
Q_2	0	1	1	1	1	2	2
Q_3	0	1	2	2	2	2	2
Q_4	0	1	1	2	2	3	3
Q_5	0	1	2	2	2	3	3
Q_6	0	1	2	2	3	3	4
Q_7	0	1	2	2	3	4	4

Fig.6. A structure of the matching of similar scenes

where $\omega_1=\omega_2=0.3$, and the rest weights are set as 0.1. If the result is larger than the empirical threshold, the two representative frames are judged as “similar”.

Scene-Level Similarity Matching. The scene-level similarity matching is realized as shown in Fig.6, where Q_i and T_j refers to the i^{th} and j^{th} representative frame in scene V_q and V_t respectively. If Q_i is similar to T_j , then $Q_i=T_j$. The table shown as Fig.6 is constructed by the equation below:

$$c[i][j] = \begin{cases} c[i-1][j-1]+1 & \text{If } Q_i = T_j \\ \max(c[i-1][j], c[i][j-1]) & \text{If } Q_i \neq T_j \end{cases}$$

It can be seen that, if the number of representative frames of V_q, V_t is M, N respectively, then the max matching frame number is $c[M][N]$. If $c[M][N]$ is larger than the empirical threshold, then V_q is judged as similar to V_t .

Video Similarity Matching. The principle of video matching is the same as that of scene-level matching. Here, Q, T_j stands for the $i^{\text{th}}, j^{\text{th}}$ scene of video V_q, V_t respectively.

3 Experimental Results

To validate the performance of the proposed near-duplicate video detection method, 50 video sequences from the TRECIVD2011 video database are tested. All the sequences are processed with 7 types of attack, including blur, contrast, crop, picture-in-picture (PIP), ratio, shift and white noise, generating a total of 350 video sequences. Recall and Precision[15] is selected in this paper to evaluate the performance of the proposed method.

A series of representative frames are shown in Fig.7. It can be seen that the representative frames are clear which can well characterize the information contain in the original videos.

The experimental results of the proposed near-duplicate video detection method are shown in Table 1. It can be seen that, when applied to videos under attacks of blur, crop, ratio, shift and white noise, the proposed method provides satisfactory results. Compared to these results, there is a decline in that of the videos under PIP and contrast attacks. PIP gives the worst results, that is because the inserted picture not only affects the sub-block based color features and edge features, it will also bring great changes to the grayscale OM sequence when appear in the middle of the frame. Contrast change also dramatically affects the result, for it will bring great changes to the grayscale features.



Fig.7. Part of the representative-sequences

Table 1. Detection results for different types of attack

	Recall	Precision
Blur	84%	100%
Contrast	74%	88.1%
Crop	92%	76.7%
PIP	66%	86.8%
Ratio	100%	90.9%
Shift	96%	85.7%
White Noise	96%	100

4 Conclusions

In this paper, a near-duplicate video detection method based on representative-sequence is proposed. Firstly, the video is divided into different scenes according to the variance of the Chi-squared color histogram and Grayscale OM (Ordering Measure), and then the frames of a single scene are preprocessed by decreasing the frame rate to a particular rate and detecting the occurrence of black-sides. For each scene, every certain number of frames are merged into one TIRI frame, and then weighted grayscale mutual information is calculated to select the representative frames from the TIRI frames. Next, several features are extracted from the representative frames for matching, including the mean, standard deviation, contrast, and an OM sequence of grayscale, as well as color distribution descriptors, and an OM sequence of edge information. Finally, whether or not the video is a near-duplicate is determined by the match degree of the characteristics. Experimental results show that the algorithm presented by this paper possesses good robustness and detection capabilities.

Acknowledgments. The work in this paper is supported by the Program for New Century Excellent Talents in University (No.NCET-11-0892), the Specialized Research Fund for the Doctoral Program of Higher Education (No.20121103110017), the National Natural Science Foundation of China (No.61003289, No.61100212), and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD201304036).

References

1. Lian, S., Nikolaidis, N., Sencar, H.T.: Content-Based Video Copy Detection – A Survey. *Intelligent Multimedia Analysis for Security Applications* 282, 253–273 (2010)
2. Roopalakshmi, R., Ram Mohana Reddy, G.: Recent Trends in Content-Based Video Copy Detection. In: 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–5 (2010)
3. NaPhade, M.R., Yeung, M.M., Yeo, B.L.: Novel scheme for fast and efficient video sequence matching using compact signature. In: Proc. SPIE – Internet Imaging, vol. 3972, pp. 564–572 (2000)
4. Jain, A.K., Vailaya, A., Xiong, W.: Query by video clip. *Multimedia Systems* 7(5), 369–384 (1999)
5. Mohan, R.: Video sequence matching. In: Acoustics, Speech and Signal Processing. In: Proceedings of the 1998 IEEE International Conference, pp. 3697–3700 (1998)
6. Hoad, T.C., Zobel, J.: Detection of video sequence using compact signatures. *ACM Transactions on Information Systems* 24, 1–50 (2006)
7. Lowe, D.G.: Distinctive image features from scale-invariant key points. *International Journal of Computer Vision* 60, 91–110 (2004)
8. Heikkila, M., Pietikainen, M., Schmid, C.: Description of interest regions with local binary patterns. *Pattern Recognition* 42(3), 425–436 (2009)

9. Law-To, J., Buisson, O., Gouet-Brunet, V., Boujema, N.: Robust voting algorithm based on labels of behavior for video copy detection. In: ACM Multimedia, MM 2006 Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 835–844 (2006)
10. Lan, D.-J., Ma, Y.-F., Zhang, H.-J.: A Novel Motion-Based Representation for Video Mining. In: ICME 2003, pp. 469–472 (2003)
11. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujema, N., Stentiford, F.: Video copy detection: a comparative study. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR 2007), pp. 371–378 (2007)
12. Esmaeili, M.M., Fatourechi, M., Ward, R.K.: A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting. IEEE Transactions on Information Forensics and Security 6(1), 213–226 (2011)
13. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: multimedia content description interface. John Wiley & Sons (2003)

Non-negative Sparse Coding Using Independent Multi-Codebooks for Near-Duplicate Image Detection

Shan Zhou¹, Jun Li², Junliang Xing³, Weiming Hu³, and Jinfeng Yang¹

¹College of Aviation Automation, Civil Aviation University of China, Tianjin, China
ss_zhou@yeah.net, jfyang@cauc.edu.cn

²School of Automation, Southeast University, Nanjing, China
lijun_automation@seu.edu.cn

³Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jlxing,wmhu}@nlpr.ia.ac.cn

Abstract. In this paper, we propose an efficient approach for detecting near-duplicate images and make three contributions as follows. First, for each sub-region of spatial pyramid, we learn one distinct codebook such that independent multi-codebooks (IMC) are produced. IMC is more accurate than traditional codebook because it considers the spatial information of visual words to a certain extent. Second, we adopt non-negative sparse coding (NSC) technique to encode features. This encoding scheme can effectively encourage similar features to share similar sparse representations. Third, we design an improved intersection kernel (IIK) to compute image similarity. We validate our approach on two datasets respectively, namely our 6K dataset where images are collected from three web image search engines and publicly available University of Kentucky dataset. The experimental results demonstrate our technique achieves significant performance gain compared with state-of-the-art approaches.

Keywords: Near-duplicate image detection, multi-codebooks, non-negative sparse coding, improved intersection kernel.

1 Introduction

With the rapid growth of network and multimedia technology, many near-duplicates or variants of the same image is widespread on the web. Given a query image, the results retrieved from current image search engines, such as Google, Baidu, Bing, often contains duplicate versions. It is necessary to identify these near-duplicate images and remove them for some applications, e.g. image retrieval and copyright detection. We address this problem in this paper, referred to as Near-Duplicate Image Detection (NDID), to find all the near-duplicates of an image on the web.

Zhang and Chang [1] define three categories of near-duplicate images, i.e. scene, camera and image. In this paper, we mainly concentrate on the near-duplicate images of type *image*, where we assume that the original image and its near-copy version share the same digital source. We also investigate the detection of near-duplicate images containing minor camera-type transformation. Fig. 1 shows some representative examples of near-duplicate images.

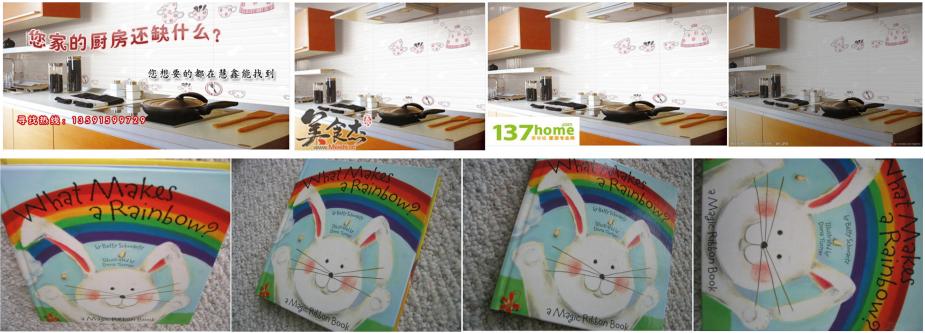


Fig. 1. Examples of duplicate images from web (the top row) and from University of Kentucky dataset [13] (the bottom row)

The remainder of the paper is organized as follows. Section 2 reviews the related work regarding NDID. Section 3 describes our framework and gives specific details of each step. Section 4 presents the experiments implemented on the two datasets and the analysis of the experimental results. Finally, in section 5, we draw a conclusion.

2 Related Work

Recently, some researchers applied the framework of image classification and retrieval to NDID for its close connection to them [2, 3]. To address the problem, bag-of-words (BoW) [4] model is considered to be an efficient method and extensively applied to NDID. The BoW-based way produces the global histogram representation of an image by encoding its local features and calculates the similarity between two global representations. It generally includes the following steps:

- (1) Extraction of local features from images, e.g. SIFT and PHOG.
- (2) Codebook training via unsupervised learning, e.g. K-means and GMM.
- (3) Encoding and pooling of local features for generating image representation.
- (4) Computation of similarity metric between images or performing classification.

However, BoW ignores the spatial information of features. According to this limitation, Lazebnik *et al* [5] proposed spatial pyramid matching (SPM) that partitioned image into increasingly finer spatial sub-regions and computed histograms of local features for each sub-region. With sparse modeling successfully applied to image and video denoising, segmentation, super-resolution, Yang *et al* [6] developed an extension of the SPM by generalizing vector quantization to sparse coding followed by multi-scale spatial max pooling. Inspired by this, Wang *et al* [7] proposed a locality-constrained linear coding (LLC) scheme that utilized locality to constrain the sparse coding process which is more computationally efficient. The two methods achieved state-of-the-art image classification performances on several benchmarks. However, both of the two schemes have the following limitations:

- (1) All sub-regions of spatial pyramid share a unique codebook, which does not take into account the spatial variability of visual words. It can not describe the local details of an image with only one codebook utilized for every sub-region of spatial pyramid.

(2) The encoding scheme does not have enough constraints to coefficient. Negative coefficients are required to satisfy the sparse coding constraints and expected reconstruction error. While adopting max-pooling on spatial pyramid constructed for an image, the information loss for the negative coefficients is inevitable.

About similarity metric for comparing distributions of features, Euclidean metric is a conventional one. However, Maji *et al* [11] have shown that Euclidean distance is not the most effective way for comparing two histograms. They built a nonlinear intersection kernel (IK) based on histogram representation and reported superior results for image classification. Thus, IK attracts much attention for low computational complexity. Nevertheless, IK just considers the minimum value of pair-wise.

According to the above analysis, we make some improvement on codebook learning and encoding constraints, and design an enhanced intersection kernel function as image similarity metric. The novel framework is applied to NDID and achieves promising performances.

3 Overview of Our Framework

In this paper, we propose a NDID system where the BoW model is combined with spatial pyramid. In training phase, we learn one codebook offline corresponding to each sub-region such that independent multi-codebooks (IMC) are produced. Then we develop a spatial pyramid image representation based on non-negative sparse coding (NSC) of low-level descriptors. Next we design an improved intersection kernel function (IIK) as the similarity metric. We finally conduct comparative experiments on our 6K dataset and University of Kentucky dataset for each phase. Fig. 2 shows the flowchart of our method. The details of each part will be elaborated as follows.

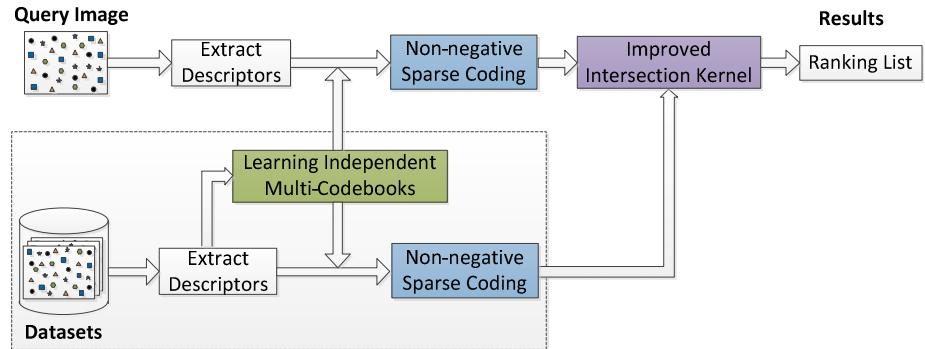


Fig. 2. The flowchart of our proposed method

3.1 Learning Independent Multi-Codebooks

Spatial pyramid partitions an image into increasingly finer spatial sub-regions and computes the corresponding BoW histogram built on a pre-trained codebook for each sub-region [8]. To our knowledge, all the state-of-the-art methods to date just train a unique codebook for the spatial pyramid, which discards the local spatial variability

of the codebook. Thus this unique codebook simply represents a coarse distribution of local features and not fully fitted to the descriptors from different sub-regions of spatial pyramid. So we introduce the idea of leaning independent multi-codebooks (IMC). For each sub-region per level of spatial pyramid, we learn a distinct codebook using corresponding local features. Once we have obtained a set of codebooks, we are able to encode the local features employing each independent codebook. Fig. 3 illustrates the comparison between one codebook and IMC. It is observed that IMC considers more spatial information and is more robust than one codebook.

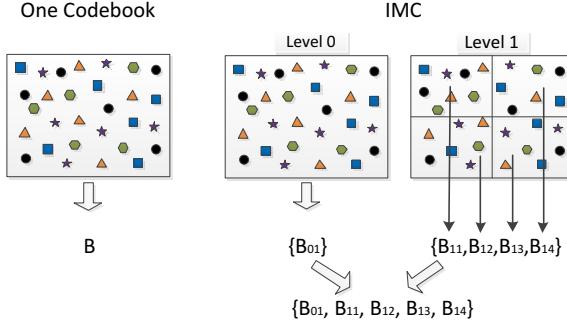


Fig. 3. Comparison between one codebook and IMC

In our experiment, we utilize the iterative online dictionary learning [9] to learn and optimize each independent codebook respectively, which construct multi-codebooks. It processes one sample at a time and sequentially minimizes a quadratic local surrogate of the expected cost. This codebook learning method can decrease the time cost in the training phase in an offline fashion to a great extent.

3.2 Non-negative Sparse Coding

Let $X = [x_1, \dots, x_K] \in R^{D \times K}$ be a set of local features in D dimensional space. Given a codebook $B = [b_1, b_2, \dots, b_M] \in R^{D \times M}$ containing M visual words of dimensionality D , the encoding technique using vector quantization (VQ) can be expressed as follows:

$$\arg \min \sum_{i=1}^K \|x_i - B\alpha_i\|^2 \quad s.t. \quad \|\alpha_i\|_0 = 1, \|\alpha_i\|_1 = 1, \alpha_i > 0. \quad (1)$$

From (1), we can observe VQ results in large quantization error, since each local descriptor is only assigned to its nearest neighbor of the codebook. Yang *et al* [6] relaxed the constraint on α_i , which enforced α_i to have a small number of nonzero elements. It is also referred to as sparse coding (SC):

$$\arg \min \sum_{i=1}^K \|x_i - B\alpha_i\|^2 + \lambda \|\alpha_i\|_1, \quad (2)$$

where λ is a regularization parameter constraining the sparsity of α_i . SC achieves good performances on several benchmarks when only SIFT descriptors are used.

Nevertheless, SC has one significant disadvantage which consists in the influence of max-pooling on the negative encoding coefficients, namely the max-pooling of local descriptors may lead to the removal of any negative coefficients with at least zero terms preserved. Hence, this strategy will cause significant information loss and the resulting image representation is not discriminative enough.

In this work, we propose an alternative non-negative sparse coding (NSC) which can be formulated as the following optimization problem:

$$\arg \min \sum_{i=1}^K \|x_i - B\alpha_i\|^2 + \lambda \|\alpha_i\|_1 \quad s.t. \quad \forall i \quad \alpha_i \geq 0. \quad (3)$$

It is can be observed that there are non-negative constraints on the coefficients. Therefore, NSC will overcome the aforementioned problem to a certain extent. Simultaneously, it can effectively encourage similar descriptors to share similar sparse representations. The problem described by (3) is generally termed the least absolute shrinkage and selection operator (LASSO). We adopt the homotopy-LARS algorithm [10] which has been demonstrated to be very efficient for solving (3).

3.3 Improved Intersection Kernel for Measuring Image Similarity

Let $H_1 = [H_{11}, H_{12}, \dots, H_{1W}] \in R_+^W$ and $H_2 = [H_{21}, H_{22}, \dots, H_{2W}] \in R_+^W$ be two histogram representations. The intersection kernel of two histograms for similarity measure is defined as follows:

$$Sim(H_1, H_2) = \sum_{i=1}^W \min(H_{1i}, H_{2i}). \quad (4)$$

This expression only takes into account the minimum value of pair-wise components without considering maximum one which is also discriminative characteristics. Motivated by this, we propose an improved intersection kernel (IIK) function described as follows:

$$Sim(H_1, H_2) = \sum_{i=1}^W (\max(H_{1i}, H_{2i}) - \min(H_{1i}, H_{2i})). \quad (5)$$

If the two histograms are similar, the difference between the maximum and the minimum values of their pair-wise components is relatively small. Compared with its prototype, IIK can capture more salient differences between two histogram sequences and quantify the histogram distances more accurately. Moreover, it is robust to the noise and variance, since it is a finer expression of the distance of two histograms. Compared with Euclidean distance and Chi-square, IIK has lower computational complexity and the optimal real-time performance.

4 Experimental Implementation and Results

In order to validate the performance of our proposed NDID method, we conduct experiments on two datasets: manually collected 6K web-image dataset, and University of Kentucky dataset [13]. We also carry out the comparative study on the

two datasets for other methods. In preprocessing stage, all images are transformed into gray-scale versions. We just use a single descriptor SURF [12], which approximates or even outperforms SIFT descriptor in terms of computational efficiency and robustness. Moreover, the dimensionality of a SURF descriptor is 64, which is merely half of a SIFT.

4.1 Experimental Settings

Since the setup of spatial pyramid will significantly influence the dimensionality of the final concatenated vector. In order to compromise between computational efficiency and precision, we set the level $L=1$, and choose $2^l \cdot 2^l$ sub-regions, i.e. the number of sub-regions overall is $N = \sum_{l=0}^L 4^l = \frac{1}{3}(4^{L+1} - 1) = 5$. Therefore, the set of multi-codebooks contain five independent codebooks learned from training local descriptors of each sub-region. We also make use of max-pooling strategy for producing the final global representation as in [7]. The pooled features from each sub-region are concatenated and normalized as the final image histogram representation. Here, we use l_2 normalization method.

4.2 6K Dataset

The 6K web-image dataset is collected from three web image search engines (Google, Baidu, and Sogou). This dataset consists of 8 class objects, i.e. animal, landmark, logo, man, musical instrument, plant, scene and transport respectively. For each class, we manually choose 25 different seed images. Therefore, there are 200 seed images overall. For every seed image, we artificially select 30 near-duplicate images from the results retrieved from the three search engines. If the search engines return less than thirty results, we transform the seed image by using ImageMagick [14], to make up to thirty near-duplicates.

Taking into account the storage and computing cost, all images are resized to be not more than 500×500 pixels. We choose 5 near-duplicate images for each seed image as training images. The low-level SURF descriptors of the training images extracted from each separate sub-region are used to train multi-codebooks. The size of each independent codebook is empirically set to 400. The precision is measured in terms of the number of relevant images in the retrieved top 30 images. We take the average precision for each class and all the query images as our evaluation criterion.

We perform three sets of comparative experiments in terms of learning different codebook, encoding methods and distance metrics on this dataset. First, we compare the effects of using IMC for our method, and the results are shown in italic in the last column of Table 1, which demonstrates the performance gain by using IMC. Next, we validate three state-of-the-art encoding methods: the baseline hard VQ, locality-constrained linear coding (LLC) [7] and our NSC. As shown in bold in Table 1, NSC outperforms others for all but one class. Last, to verify the effectiveness of IIK for our approach, we compare other three similarity metrics, i.e. Euclidean distance, Chi-square distance, intersection kernel (IK), and the results are shown in italic in Table 2. It is explicitly observed that our IIK achieves the optimal performance.

Table 1. The precision (%) between different encoding schemes for our method, the last column in italic shows the result using IMC on NSC

		VQ	LLC	NSC	IMC+NSC
object class	animal	93.20	95.60	96.13	96.93
	landmark	89.20	93.60	94.93	95.47
	logo	94.13	98.27	98.40	97.73
	man	84.00	95.07	94.80	95.20
	musical	88.00	94.53	94.80	95.73
	plant	92.80	95.47	96.66	97.33
	scene	89.20	95.60	96.80	97.60
	transport	88.80	96.93	97.60	98.00
	Average Precision (%)	89.92	95.63	96.26	96.75

Table 2. The precision (%) between different similarity metrics for our method

	Euclidean	Chi-square	IK	IJK
Precision (%)	93.82	94.82	83.50	96.75

4.3 University of Kentucky Dataset

University of Kentucky (UK) dataset includes 10, 200 images of 2550 objects where each object has 4 relevant images with different camera viewpoints or angles. The size of all the images in this database is 640×480 pixels. We resize all images to 480×360 pixels. We uniformly sample feature descriptors from all images to train multi-codebook. Every independent codebook of multi-codebooks is trained with 1000 bases. Our evaluation criterion is to calculate an average over the number of true positives of returned top four images when using a query image randomly chosen from that set of four images.

We also conduct three sets of comparative experiments on this public dataset, and the corresponding results are shown in Table 3, Table 4. It can be observed that the experimental results may be outperformed by the ones reported in some related literature, since the four relevant images of one object has different camera viewpoint whereas spatial pyramid technique is not robust enough to the variance of camera viewpoint. In spite of this, our proposed method outperforms other approaches overall.

Table 3. The average top between different encoding schemes for our method, the last column in italic shows the result using IMC on NSC

	VQ	LLC	NSC	IMC+NSC
Average Top	2.7055	2.7702	2.9733	<i>3.0012</i>

Table 4. The precision (%) between different similarity metrics for our method

	Euclidean	Chi-square	IK	IJK
Average Top	2.6980	2.9824	1.3212	3.0012

5 Conclusion

This paper presents an improved framework for near-duplicate images detection by combining the BoW model with spatial pyramid. This framework mainly consists of

three elements, namely independent multi-codebooks, non-negative sparse coding and improved intersection kernel function (IMC+NSC+IIK). We learn IMC, which consider more spatial information of visual word to an extent. In addition, we adopt NSC to encode the low-level descriptors with lower information loss. We also design an IIK function to measure the similarity between two images. Experimental results on two datasets validate the proposed three improved parts effectively enhances the detection performance of near-duplicate images.

Acknowledgement. This work is partly supported by NSFC (Grant No. 60935002 and 61273023), the Ph.D. programs Foundation of Ministry of Education of China (Grant No. 20120092110024), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Zhang, D., Chang, S.F.: Detecting Image Near-duplicate by Stochastic Attributed Relational Graph Matching with Learning. In: ACM Multimedia Conference, pp. 877–884 (2004)
2. Dong, W., Wang, Z., Charikar, M., Li, K.: High-Confidence Near-Duplicate Image Detection. In: ACM International Conference on Multimedia Retrieval (2012)
3. Meng, Y., Chang, E.Y., Li, B.: Enhancing DPF for near-Replica Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. II-416-23 (2003)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: 8th European Conference on Conference Vision, pp. 1–22 (2004)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
6. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear Spatial Pyramid Matching using Sparse Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)
7. Wang, J., Yang, J., Fu, K., Lv, F., Huang, T., Gong, Y.: Locality-Constrained Linear Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
8. Gauman, K., Darrell, T.: The Pyramid Match Kernels: Discriminative Classification with Sets of Image Features. In: Conf. Comput. Vision Pattern Recognit., pp. 1458–1465 (2005)
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. *J. Mach. Learn. Res.* 19–60 (2010)
10. SPArse Modeling Software,
<http://spams-devel.gforge.inria.fr/index.html>
11. Maji, S., Berg, A.C., Malik, J.: Classification Using Intersection Kernel Support Vector Machines is Efficient. In: IEEE Conf. Comput. Vision Pattern Recognit., pp. 1–8 (2008)
12. 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)
13. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
14. ImageMagick, <http://www.imagemagick.org/>

Machine Vision Based Automatic Micro-parts Detection System

Xianchuan Yu, Guanyin Gao, Wu He, and Jindong Xu

College of Information Science and Technology, Beijing Normal University, Beijing 100875

Abstract. Traditional micro-parts defect detection methods are based on human observation to find parts defects, which is inefficient and inaccurate. In order to solve these problems, we designed an automatic parts defect detection system based on machine vision, which include parts transfer module, image capture module, image recognition module and parts selection module. Besides, we put forward a method which is suitable for micro-parts. Experiments show that the method works well and can achieve high accuracy as well as high efficiency, which helps reduce the cost of labor, improve efficiency, production quality and automation.

Keywords: Machine vision, parts detection, image processing, template matching.

1 Introductions

Micro-parts are parts with their size between 0.01mm and 10mm, namely micro/meso scale parts [1]. Micro-parts have small sizes, complicated shapes and high precisions, traditional manual detection can't meet current needs, and thus it's an extremely urgent task to develop an automatic detection system of high precision and efficiency. Current researches only focus on the aspect of visual parts inspection, neglecting the automatic sifting [2]. This article introduces an accurate and efficient automatic detection system aimed at micro-parts sifting.

2 Overall Design Scheme

As shown in Fig1, the overall design scheme of our system is composed of the parts transfer module which consists of a round glass turntable and the electric motor, image capture module which includes the illuminant, CCD camera, IO card for control signal transmission and image acquisition card, image processing module which mainly works on PC and the parts sifting module which includes pinhole air blow tank, etc. [3,4].

The system works as follow. The glass turntable drives the parts rotating at constant velocity, when parts enter the view field of CCD camera. Images are taken and transferred to PC by the image acquisition card. PC processes the images, including a

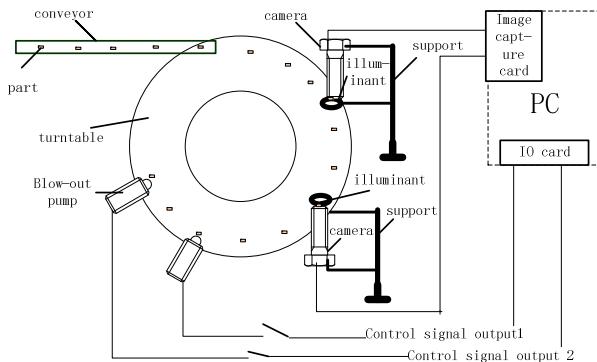


Fig. 1. System overall design scheme

pre-processing step for interesting region, a classification step using pattern recognition methods. Control signals are transferred to the parts sifting module by Data Acquisition Card. CCD camera is controlled by software [5], and the interval of photographing is set with respond to the rotating velocity and the parts interval, only to ensure the synchronization of camera and parts transfer module. If a part is recognized as unqualified and needs to be removed, the IO card sending the removing signal to the parts sifting module T seconds later.

3 System Hardware Design

3.1 Design of Parts Transfer Module

The targeting parts of our system are irregular white cuboids. A transparent glass turntable is driven by the electrical motor, which rotates in a constant velocity. Parts slide onto the conveyor belt at a certain time interval, with the machinery open and close periodically. Then the parts are conveyed to the glass turntable, where parts rotate in a constant velocity and keep a uniform interval.

Both sides of the parts should be detected. Therefore, we capture two images for each side. CCD cameras are settled at both the upper side and the underside, thus enabling system to take images without flipping the parts. Moreover, the center of the glass turntable is an empty circle, which allows the removed parts fall down conveniently.

3.2 Design of Image Capturing Module

It's critical for machine vision system to classify and detect parts to capture high quality images. When designing the illuminant, we chose the RL-96-00 24V low angle annular red kind, considering the camera view field, distance between illuminant and parts, shape of the parts, color, illumination, etc. [6].

Pike-F505B version of CCD camera is used for its high accuracy, stability and anti-vibration, whose Optical format is 2/3 inch. Two staggered sets of image acquisition unit with black ground are used to capture the upper-side and under-side of parts.

System chooses DH-CG300 version of image acquisition card, which uses the PCI bus and hardly takes CPU time. Images are stored in static memory using APIs of the image acquisition card.

3.3 Design of Parts Removal Module

An air blow tank is used by the parts removal unit. Its small wind range ensures only one part is blew away without influencing others. The air blow tank is connected to the DAQ Card (Data Acquisition Card), thus PC transmits control signals to the DAQ to control the air blow tank indirectly.

4 Design of System Software

4.1 Flow Chart of System Main Routine

The original parts images are processed through the pre-processing step and pattern recognition step by main routine to distinguish defects. Preprocessing step includes binaryzation, denoising, image locating, rotating and translation. Pattern recognition step includes the template based edge detection and pits recognition on the surface. Fig 2 shows the main routine of detection flow chart.

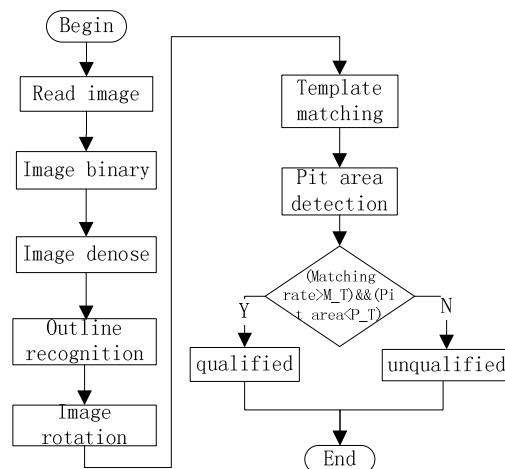


Fig. 2. System main routine of flow chart

4.2 Image Preprocessing

4.2.1 Image Binaryzation

Set the threshold as 120. The pixel value which is less than the threshold is set to be 0, the pixel value which is greater than the threshold is set to be 255

4.2.2 Image Denoising

After image binaryzation step, some isolated points come into being. Common denoising algorithms aim at gray image denoising, which may not be suitable for binary images. We put forward a new median filtering based image denoising algorithm, aiming at the noisy points in binary images.

Our denoising algorithm works as follow. Scan the whole image from lower left corner line-by-line, if we encounter a white point, calculate the ratio of white points versus black points in its 5×5 neighborhood. A white point is recognized as noisy if the ratio is lower than $3/5$, thus setting its value to be 0 and go on with the next point. Otherwise, it's a normal point and we don't operate on it.

4.2.3 Outline Recognition and Normalization

We take the lower left corner of image as point $(0, 0)$, and the first task is to find out the four corner coordinates. The abscissa of every point is its row number and the ordinate for every point is its column number. Firstly, we scan the image line by line from left upper corner, and the first white point is regarded as the upper most coordinate, thus setting its height to 1, go on with the next line. When there are at least one white point in a line, increase height and go on. Otherwise, compare height with half of width of part. If height is greater than half of width of part, we can determine the upper most coordinate and exit the loop. Else, we can see the first white point as noise point and set its pixel to 0. Scanning the image again until we determine the upper most coordinate. The coordinates of the other three corners are determined by the same method. After the four corners coordinates are determined, we can achieve the center point coordinate of part. The abscissa of the center point is the mid-value of the abscissa of upper most corner and the abscissa of lower most corner. The ordinate of the center point is the mid-value of the ordinate of left most corner and the ordinate of right most corner.

4.2.4 Image Rotation and Translation

Image rotation is the process of rotating image with a central point and a certain angle to form a new image. Because of the angle of each part is not the same, the angle of inclination should be calculated.

The Least Square Method is used to calculate the angle of inclination. The lower most and right most coordination determines a line, and the Least Square Method uses all the white points to fit the line, then we get the slope.

$$k = \frac{N \sum_{i=1}^N X_i Y_i - (\sum_{i=1}^N X_i) (\sum_{i=1}^N Y_i)}{N \sum_{i=1}^N X_i^2 - (\sum_{i=1}^N X_i)^2} \quad (1)$$

k means the slope, N mean the number of white points, X_i is the abscissa of white points, Y_i is the ordinate of white points.

Since the center of image and distance to center never changes after rotation, we can get the corresponding relationship between point coordination after and before.

In the new coordinate system, the distance between (x_0, y_0) and the original point is r , and the angle of the line composed by these two points between x axis is b , the rotation angle is a , thus we get the corresponding point (x_1, y_1) after rotation, shown as Fig 3.

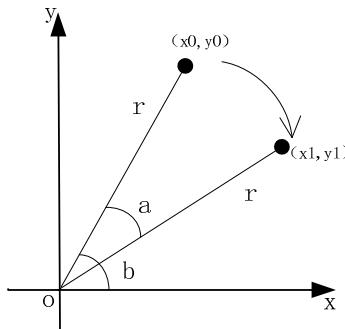


Fig. 3. Sketch map of rotation

$$\begin{cases} x_0 = r \cos b \\ y_0 = r \sin b \end{cases} \quad (2)$$

$$\begin{cases} x_1 = r \cos(b-a) = r \cos b \cos a + r \sin b \sin a = x_0 \cos a + y_0 \sin a \\ y_1 = r \sin(b-a) = r \sin b \cos a - r \cos b \sin a = -x_0 \sin a + y_0 \cos a \end{cases} \quad (3)$$

Then we need to translates the parts portion of image to the image center. Let's assume the position of parts center to be (x_l, y_l) , the position of image center to be (x_t, y_t) , translates process as shown in formula (4).

$$\begin{cases} x_1 = x_0 + (x_t - x_l) \\ y_1 = y_0 + (y_t - y_l) \end{cases} \quad (4)$$

4.3 Recognition and Adjudication

4.3.1 Template Matching to Detect Part Edges

Now, there are many template matching algorithms that have been put forward, among which relevance based matching and shape matching are the mainstream. We adopt the parts outline based matching algorithm according to the possible edge defects such as saw tooth, unfilled corner, and pits. In other words, comparisons are made between corresponding points of preprocessed image and the template image, which contains the standard parts outline.

Scan the parts image. Calculate length of the four edges according to the coordinates of four part corners. Then recognize the status of rotated parts, horizontal or vertical. If it's horizontal, horizontal template image is used for matching, otherwise, vertical one is used.

We scan the parts image and template image simultaneously from the upper left corner, and count the number of white points that matched between them, which is the number of points that part image matched the template image. Match ratio is calculated according to the number of matched points and number of white points. Finally, adjudicate whether or not the detected part has defects or not.

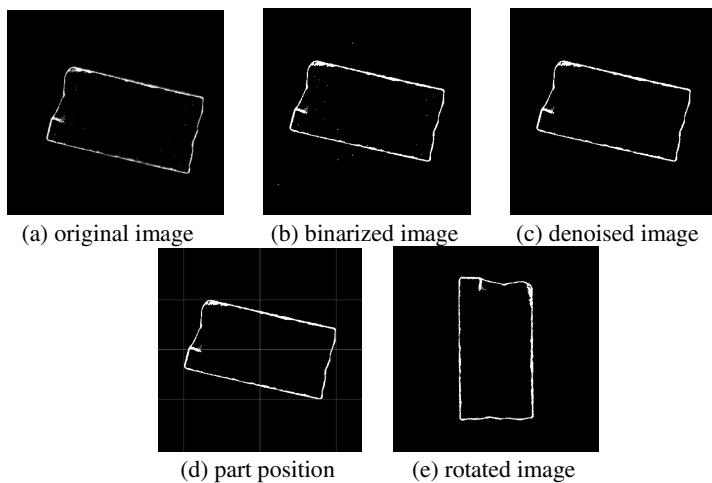
4.2.2 Pits Detection on Parts Surface

There will be inevitably small pits on the surface, and too much pits would make the part unqualified. Firstly, we determine the inner edge of part by scanning row by row or column by column.

After having confirmed the four inner edges, for every point calculate the number of white points in the 5×5 neighborhood. When we have finished the process for every point, find out the maximum number of white points in the 5×5 neighborhood. If the number excels system threshold, it means that the current part unqualified. The threshold can be set to be 20.

5 Results

The detection software of this system is developed in the compiler environment of visual studio 2008. It is tested in the machine with the configuration is Pentium(R)D processor(CPU is 2.5GHZ), taking time less than 250ms, so that this system can achieve real-time processing for image and the classification accuracy is greater than 90%.

**Fig. 4.** Image preprocessing

This part is defective. Fig4a displays the original image. Fig4b is binarized image. In this image, there are some isolated points. Fig4c is denoised image. Most of the isolated points are removed. Fig4d is the part area is determined by the coordinates of the four corners of the part. Fig4e is the rotated image. In this image, part is rotated to vertical state. The match rate of this image is 0.29, and the maximum pit area is 25 pixels.

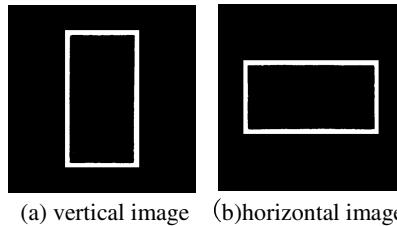
**Fig. 5.** Template image

Fig5 displays two template images. For increasing the fault-tolerant space, the contour of part in template image is bold.

In order to test the accuracy of the detection system, 200 qualified part samples and 300 unqualified part samples were selected for testing.

Table 1. Distribution of qualified and unqualified parts matching rate

Num of part Matching rate	Selecting 200 qualified parts	Selecting 300 unqualified parts
0.8~1.0	173	42
0.6~0.8	21	77
below 0.6	6	181

It can be seen from the table1 that qualified and unqualified parts matching rate distribution have evident differences. More than 85% qualified parts matching rate are above 0.8, only a few qualified parts matching rate below 0.6; the majority of unqualified parts are distributed in the matching rate below 0.6. A good classification effect can be achieved, if choosing a matching rate threshold between 0.6 and 0.8.

Table 2. Distribution of qualified and unqualified parts pit area size

Num of part Pit area (Pixel)	Selecting 200 qualified parts	Selecting 300 unqualified parts
20~25	4	166
10~20	38	89
below10	158	45

It can be seen from the table2 that the pit area of most qualified parts are below 10 pixel, and only a few qualified parts pits area are above 20 pixels; The pit area of unqualified parts is generally larger, mostly distributed above 20 pixels. The differences of parts pit area distribution is obvious.

6 Conclusion

In order to resolve the practical problems of improving the micro-parts detection accuracy and detection efficiency, this paper researched the existing automatic detection technology and parts detection equipment, developed an automatic detection device for micro-part based on machine vision, and developed the defect detection software. The experimental results show that: the system can achieve automatic detection, improving the detection accuracy and efficiency.

References

1. WORKSHOP R. Workshop on micro/meso-mechanical manufacturing, Evanston, Illinois, USA, Northwestern University (2000)
2. Li, B.-H., Wang, S.-L., Li, X.-B.: An automatic gauging system for crankshaft. Mechanical Science and Technology for Aerospace Engineering 28(4), 510–513 (2009)
3. Geng, C.-M., Cai, D.-B.: Design of image measurement system for mechanical parts based on computer vision. Computer Measurement & Control 20(1), 38–40 (2012)
4. Wei, S., Quan, P., Ying-Le, F., Ping, X.: Automatic automobile parts recognition and classification system based on machine vision. Instrument Technique and Sensor (9), 97–100 (2009)
5. Jiang, G.-W., Chao, Z.-C., Jiang, H.-P., Fu, S.-H.: Synchronous image acquisition and processing system of multiple cameras based on source trigger and software control. Journal of Applied Optics 30(5), 756–760 (2009)
6. Wang, Y.-T., Yang, N., Lin, X.-L.: Research on float glass thickness measurement system based on CCD technology. Instrument Technique and Sensor (6), 72–74 (2009)

Segment Based Depth Extraction Approach for Monocular Image with Linear Perspective

Yiming Mo¹, Tianliang Liu^{1,*}, Xiuchang Zhu¹, Xiubin Dai¹, and Jiebo Luo²

¹ Nanjing University of Posts and Telecommunications, Nanjing, 210003, China
moyimingmm@gmail.com, {liutl, zhuxc, daixb}@njupt.edu.cn

² University of Rochester, NY 14627, USA
jluo@cs.rochester.edu

Abstract. In this paper, a segment-guided depth extraction approach is proposed for monocular image with linear perspective. Firstly, foreground depth is learned from a RGBD database with segment-based calibration to adjust the initial coarse depth, and background depth is estimated from linear perspective by vanishing cues. Then, the foreground depth and background one are linearly combined with a statistically optimal balance factor to obtain a holistic fused depth map. Lastly, bilateral filter is exploited to suppress the depth disturbance with edge-preserving. Experiments demonstrate that the proposed technique can produce accurate and dense depths with distinct object boundaries and correct relation among the object positions for a single image.

Keywords: depth extraction, monocular image, segment-guided calibration, linear perspective, foreground/background.

1 Introduction

Depth estimation is one of the most fundamental problems in 2D to 3D conversion of monocular image. Previous works on depth extraction focus on the traditional strategies, such as structure from motion [1] and depth from defocus [2]. However, these estimation methods depend to a very large extent on camera parameters (pose and focal length), which are difficult or even impossible to obtain in some cases.

Recently, data-driven machine learning techniques are put forward to estimate depth map by Saxena et al. [3], and Karsch et al. [4] extend Saxena’s learning procedure by transferring depth from video with a large repository of RGBD database. Um et al. [5] demonstrate that depth estimation with image segments can greatly improve visual effects of depth results. Meanwhile, geometry cues can reflect depth distribution from other aspects. For example, Lai et al. [6] apply vanishing information to depth extraction by utilizing a principal linear perspective.

In this work, we cast the problem of depth extraction by fusing foreground depth and background one to obtain a holistic depth map with statistical opacity balance

* Corresponding author.

factor and bilateral filter. Foreground depth is acquired by data-driven machine learning and segment-based amendment, while the idea of linear perspective is extended to get the background by vanishing cues. Without any detailed camera parameters, the proposed approach is still able to obtain high quality depth with obvious structure, distinct object boundaries and accurate relative scene positions.

2 Proposed Depth Estimation Scheme

The proposed approach is built upon a key hypothesis that the depth of monocular image can be composed by two parts, foreground depth that focuses on the depth of salient regions and background depth that globally reflects the overall trend of depth distribution. By combining the two parts, the holistic depth can not only demonstrate a relatively correct depth distribution tendency, but also exhibit unambiguous details. The pipeline of the whole depth extraction can be exhibited in Fig.1.

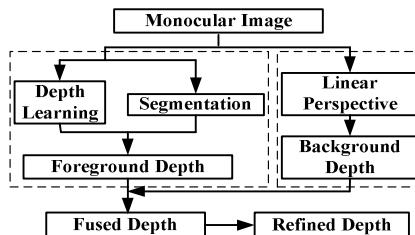


Fig. 1. Overview of depth extraction

2.1 Initial Foreground Depth Learning

Foreground depth extraction shown in Fig.2, contains two sub-procedures, initial depth learning and segment-based depth amendment. Since the visual scenes with similar semantics or photometric contents are likely to have analogous depths, three steps are consequently advised in depth learning similar to the Karsch's works [4].

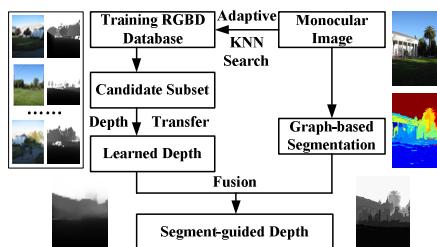


Fig. 2. Foreground depth workflow

Firstly, an adaptive K -nearest neighbor (KNN) searching is employed to extract proper candidate subsets. Then, the SIFT flow warping [10] is introduced for dense

scene alignment. Lastly, a non-parametric sampling energy function for foreground region is defined to learn an initial depth map of the input scene image.

Candidate subset containing K matching images and corresponding depths is searched by means of high level image features which reflect global information. Histogram of oriented gradient (HOG) [8] descriptor is preferred rather than “gist” feature [9]. Though “gist” features reflect global characteristics in a relatively low-dimensional scene representation, HOG descriptors are more superior in maintaining geometric consistency and photometric consistency. Fig.3 shows that HOG features are better than “gist” ones in finding the candidate subsets in scene semantic level.



Fig. 3. Candidate images chosen from gist features and HOG features respectively for img-op57-p-016t000 in Make 3D database

Since each candidate image matches input image closely in scene semantic space, there should be comparably similar depth distributions when they are densely aligned. Warping functions $\phi_i, i \in \{1, \dots, K\}$ for each candidate image is estimated through SIFT flow [10], and an initial learned depth is achieved by considering all of the warped candidate depths with a transfer strategy. Let L be input image and D_{le} the depth map to be inferred. We minimize a non-parametric sampling energy function as follows:

$$-\log(P(D_{le} | L)) = E(D_{le}) = E_t(D_{le}) + \lambda E_s(D_{le}) \quad (1)$$

where $E_t(D_{le})$ represents data term, $E_s(D_{le})$ denotes spatial smoothness term, and constant λ is a regulate factor ($\lambda=10$). The $E_t(D_{le})$ term denotes closeness measure between the learned depth D_{le} and each of the warped candidate depths, while the $E_s(D_{le})$ term encodes the depth gradients tuned by soft thresholds of image gradients.

2.2 Segment-Guided Foreground Depth

We encourage mid-level structural constraint on initial learned depth map of the input scene with suitable visual segments, which can be produced by exploiting graph-based segmentation [11]. The obtained compact segment map often consists of the expected regions satisfying global properties of appearance consistency and evident boundaries between two regions using graph-based representation of the given image. Nearest neighbor graph can be constructed rather than grid one, since the formal can provide more accurate edges of scene objects from image segmentation.

Scene segments with geometric structure constraint of input image can be projected onto the learned depth map to produces a preferable foreground depth D_f with higher quality of visual effects. Let S_j be the j th segment projected onto the learned depth

with the total number of pixels N_j . For each pixel $i \in S_j$, the depth values are averaged in the predefined segment. The depth amendment phase can be described as follows:

$$\mathbf{D}_f(i \in S_j) = \frac{1}{N_j} \sum_{i \in S_j} \mathbf{D}_{le}(i) \quad (2)$$

The segment-guided foreground depth estimation for monocular image can be listed as Algorithm 1:

Algorithm 1. Segment-guided depth amendment

- 1: Learn initial depth map \mathbf{D}_{le} .
- 2: Split input image into n segments $S = \{S_1, \dots, S_j, \dots, S_n\}$.
- 3: Let $j = 1$, $D_{total} = 0$ and $N_j = 0$.
- 4: If pixel $i \in S_j$, then turn to step 5; else turn to step 6.
- 5: Update $D_{total} = D_{total} + D_{le}(i)$; $N_j = N_j + 1$.
- 6: $D_{ave} = D_{total} / N_j$, $\mathbf{D}_f(i \in S_j) = D_{ave}$, update $j = j+1$, $D_{total} = 0$ and $N_j = 0$, if $j \leq n$, return to Step 4; else if $j > n$, obtain final foreground depth \mathbf{D}_f .

2.3 Global Background Depth Estimation

Vanishing lines or vanishing points encode accurately global structural and geometric information of monocular image. Consequently, linear perspective is introduced to assign background depth with vanishing cues extracted from the given image. Firstly, the straight lines are detected by Hough transform, then, a cluster analysis with intersection point neighborhood is adopted to estimate vanishing points from the lines [12]. Finally, depth gradients are assigned in a progressive expression along vanishing directions. Since the depths in Make3D database in our experiment were collected by a 3D scanner and have a maximum range of 81m, the depth of linear perspective should also be distributed between $D_{min}(=0)$ to $D_{max}(=81)$ meters.

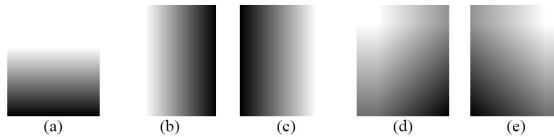


Fig. 4. Categories of linear perspective (a) up-bottom (b) left-right (c) right-left (d) upper left corner-lower right corner (e) upper right corner-lower left corner

We classify linear perspectives of the given scenes as follows:(a) up-bottom perspective, which only exists a horizontal vanishing line; (b) left-right perspective that only exists a vertical vanishing line on the left; (c) right-left perspective that only exists a vertical vanishing line on the right; (d) upper left corner-lower right corner perspective that exists both horizontal vanishing line and vertical vanishing line where the vertical vanishing line belongs to the left side; (e) upper right corner-lower left corner perspective that exists both horizontal vanishing line and vertical vanishing

line where the vertical vanishing line belongs to the right side. Fig.4 shows the five classical categories of linear perspective by normalizing D_{max} to one and D_{min} to zero.

2.4 Holistic Depth Fusion and Refinement

The foreground and background depths alone have their own strengths and weaknesses, so we combine them together to complement each other. The foreground depth \mathbf{D}_f and background depth \mathbf{D}_b are linearly fused with a reliable balance weight α through a convex combination to obtain a holistic depth map \mathbf{D}_{fuse} as follows:

$$\mathbf{D}_{fuse} = \alpha \mathbf{D}_f + (1 - \alpha) \mathbf{D}_b \quad (3)$$

where the value α denotes foreground opacity ($\alpha \in [0,1]$). This linear combination is crude since fused depth depends much on the choice of parameter α , when α is under a proper level, the holistic depth shows vague details, and if α begin to exceed a reasonable value, the holistic depth could not take full advantages of geometric cues. A statistical method is adopted to achieve a general weight that satisfies most outdoor scenes. We empirically find that this balance scheme on α works efficiently. As the Make3D range database provides ground truths as reference [3, 7], we iteratively minimize (4) and (5) simultaneously on each sample from a statistical perspective:

$$E_{max}(i) = \max(\alpha \mathbf{D}_f(i) + (1 - \alpha) \mathbf{D}_b(i) - \mathbf{D}_t(i)) \quad (4)$$

$$E_{ave}(i) = \frac{1}{N} \sum_{i=1}^N (\alpha \mathbf{D}_f(i) + (1 - \alpha) \mathbf{D}_b(i) - \mathbf{D}_t(i)) \quad (5)$$

where $i \in pixels$, N stands for total number of pixels in the assumed depth map, $\mathbf{D}_t(i)$ is the ground truth depth of pixel i . $E_{max}(i)$ represents a measure of maximum error (ME) for each pixel, and $E_{ave}(i)$ measures average error (AE) on the given depth map.

Since foreground depths are estimated from segment-guided amendment and background depths are extracted on the basis of block-level, the fused depths may not produce good visual effects. In the post processing phase, bilateral filter [13] is applied to refine depth maps. Since it evaluates the similarity of colors and distances between current pixel and its neighborhoods, the weighted bilateral filter can keep edges smooth and align depth edges with the input color edges at the same time.

3 Experimental Results

The experiments were performed on the Make3D range image dataset by A. Saxena [3, 7]. The Make3D benchmark database includes 400 training samples and 134 test samples, while each sample consists of a monocular 2D image and a corresponding depth map collected by a custom-built 3D scanner. The error threshold of HOG is set to be 0.9 in the KNN search, and the optimal balance weight α equals to 0.85 in Eq.(3), being learned from a statistical perspective. The depth evaluations are measured based on known ground truths and luminance images of input scenes.

Fig.5 compares the results of three typical monocular images from our proposed approach and state-of-the-art depth transfer [4] in visual qualities. The extracted depth maps exhibited in column (c) by our approach have distinct boundaries of objects, and more accurate object positions than that of the scene in column (b) via depth transfer [4]. Ground truths are exhibited in last column as reference depths. Table 1 shows that the qualitative results of three acquired depths above can be compared with that of the scene from depth transfer. It can be seen from these results shown in Fig.5 and Table 1 that the proposed whole depth extraction can obtain piecewise smooth, accurate and dense depth maps. The given approach may be more suitable to the visual scene with more notable global perspective cues and as distinct boundaries are mainly brought in by the segmentation process, the proposed approach may handle images with more apparent segments of the objects better other than images which contain rich edges.

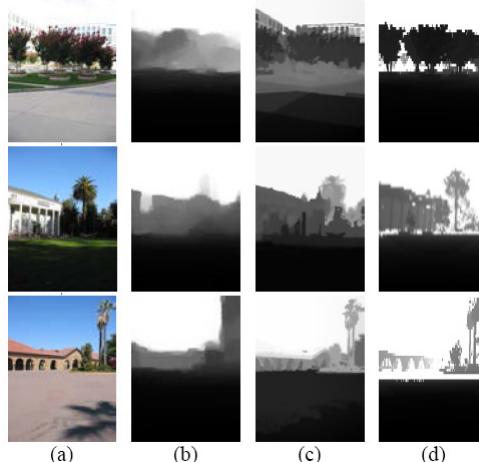


Fig. 5. Experimental results on three test images “img-op1-p-313t000”, “img-op57-p-016t000” and “img-op23-p-139t000” (from top to bottom): (a) input images, (b) results of depth transfer, (c) results of proposed approach, (d) ground truth depths

Table 1. Quantitative evaluation of the proposed method on three monocular images

Test Image	img-op1-p-313t000		img-op57-p-016t000		img-op23-p-139t000	
	Different Method	Depth Transfer	Proposed Approach	Depth Transfer	Proposed Approach	Depth Transfer
RMS(m)	19.33	17.12	15.22	14.54	21.90	16.75
SSIM	0.2928	0.4261	0.4974	0.6156	0.3456	0.5202

The quantitative measures are defined on two metrics, such as root mean squared (RMS) error and structural similarity index measure (SSIM) [14]. The RMS error quantifies the difference of the estimated depth against its ground truth. Since depth borders of the scene assumedly coincide with that of the objects in luminance with respect to structural similarity, we suggest the SSIM measure assesses the similarity

between the structures of the extracted depth map and that of the luminance of input scene, which has proven to be more consistent with human eye perception.

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{D}_o(i) - \mathbf{D}_t(i))^2} \quad (6)$$

$$SSIM(\mathbf{D}_o, \mathbf{L}) = l(\mathbf{D}_o, \mathbf{L})c(\mathbf{D}_o, \mathbf{L})s(\mathbf{D}_o, \mathbf{L}) \quad (7)$$

where \mathbf{D}_o and \mathbf{D}_t represent the estimated depth and its ground truth respectively, \mathbf{L} be luminance image of input scene. The RMS measure is evaluated in meters in the metric system, which reflects the relative errors between the estimated depth and its ground truth. The first term $l(\mathbf{D}_o, \mathbf{L})$ in Eq. (7) is luminance comparison function which measures the mean luminance closeness between the obtained depth map and gray component of the input image. The second term $c(\mathbf{D}_o, \mathbf{L})$ is contrast comparison function which is measured by the standard deviation. The third term $s(\mathbf{D}_o, \mathbf{L})$ is structure comparison function which measures correlation coefficient between the depth \mathbf{D}_o and the image \mathbf{L} . The positive values of the SSIM measure are in the range of [0, 1], while the closer to one is this numerical value, the geometric structure of the estimated depth \mathbf{D}_o is more similar to that of luminance image \mathbf{L} of input visual scene.

Table 2. Comparison of Depth Estimation Errors on the Make3D Range Image Dataset

Measure \ Method	Depth Transfer	Proposed Approach
RMS(m)	15.14	17.48
SSIM	0.2413	0.6255

Although, compared with recent state-of-the-art depth extraction methods such as depth transfer [4], our results shown in Table 2 evaluated on the whole Make3D test dataset are comparable in terms of RMS measure, and the presented depth estimation approach dramatically improve the overall accuracy with regard to SSIM assessment.

4 Conclusion and Future Work

We present an automatic depth estimation approach by fusing foreground depth and background one into an ensemble depth layer for single image. Main contributions contain three parts: firstly, foreground depth is estimated by learning and graph-based segment-guided amendment; then, linear perspective from vanishing information is utilized to obtain the background representing a holistic depth distribution; last, foreground depth and background one are fused with a statistical optimal weight. Our whole approach may be more suitable to the visual scene with notable global structural perspective cues and apparent or evident segment boundaries of the objects.

Furthermore, we plan to improve the performance of the proposed depth extraction with boundary and junction characteristics of hierarchical over-segmentation for an image, while incorporating temporal information from frame sequences for 2D video.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China (Project No: 61001152, 61071091, 61071166, 31200747 and 61172118), the Natural Science Foundation of Jiangsu Province of China (Project No: BK2010523 and BK2012437), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province of China (Project No: 11KJB510012) and the Scientific Research Foundation of Nanjing University of Posts and Telecommunications under Grants NY210069, NY210073 and NY211030. We also thank Stanford 3D Reconstruction Group for their open Make3D database.

References

- Cheng, C.M., Hsu, X.A., Lai, S.H.: A Novel Structure-from-Motion Strategy for Refining Depth Map Estimation and Multi-view Synthesis in 3DTV. In: IEEE International Conference on Multimedia and Expo, pp. 944–949. IEEE press, Suntec City (2010)
- Favaro, P.: Recovering Thin Structures via Nonlocal-means Regularization with Application to Depth from Defocus. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1133–1140. IEEE press, San Francisco (2010)
- Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D Scene Structure from Single Still Image. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(5), 824–840 (2009)
- Karsch, K., Liu, C., Kang, S.B.: Depth Extraction from Video Using Non-parametric Sampling. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 775–788. Springer, Heidelberg (2012)
- Um, G.M., Bang, G., Cheong, W.S., Hur, N., Lee, S.I.: Improvement of Segment-based Depth Estimation using a Novel Segment Extraction. In: 3DTV Conf., Tampere, pp. 1–4 (2010)
- Lai, Y.K., Lai, Y.F., Chen, Y.C.: An Effective Hybrid Depth-generation Algorithm for 2D-to-3D Conversion in 3D Displays. *J. Disp. Technol.* 9(3), 154–161 (2013)
- Saxena, A., Chung, S.H., Ng, A.Y.: Learning Depth from Single Monocular Images. *J. Adv. Neural Inf. Process. Syst.* 18, 1161 (2006)
- Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893. IEEE press, San Diego (2005)
- Oliva, A., Torralba, A.: Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Prog. Brain Res.* 155, 23–36 (2006)
- Liu, C., Yuen, J., Antonio, A.B.: SIFT Flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(5), 978–994 (2011)
- Criminisi, A., Reid, I., Zisserman, A.: Efficient graph-based image segmentation. *Int. J. Comput. Vision.* 59(2), 167–181 (2004)
- Schmitt, F., Priese, L.: Vanishing Point Detection with an Intersection Point Neighborhood. In: Brlek, S., Reutenauer, C., Provençal, X. (eds.) DGCI 2009. LNCS, vol. 5810, pp. 132–143. Springer, Heidelberg (2009)
- Tian, C., Krishnan, S.: Accelerated Bilateral Filtering with Block Skipping. *IEEE Signal Process. Lett.* 20(5), 419–422 (2013)
- Hore, A., Ziou, D.: Image Quality Metrics: PSNR vs. SSIM. In: 20th International Conference on Pattern Recognition, pp. 2366–2369. IEEE press, Istanbul (2010)

A Fuzzy Mix-Prototype Clustering Algorithm for Leukemia Data Analysis

Jin Liu, Qianping Wang, Zhizhen Liang, and Wei Chen

School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, Jiangshu 221116, China

{liujincumt,davior.chen}@gmail.com
{qpwang,liang}@cumt.edu.cn

Abstract. Following our previous work which adopts hyperplanes to represent the cluster centers in the fuzzy clustering, we present a revised version of the algorithm which combines both spherical and hyper-planar cluster centers. The presented fuzzy clustering algorithm is capable to capture latent data models with both spherical and non-spherical geometric structures, we present the formulation of fuzzy objective function and derive an iterative numerical solution for minimizing the objective function. For purposes of validations and comparisons, the proposed algorithm is applied to perform microarray data analysis on Leukemia data set, the results have shown that the presented fuzzy clustering technique provides an effective alternative for microarray data analysis.

1 Introduction

Despite a huge amount of successful applications of the fuzzy c -means(FCM) and many of its variants, it has been noticed that these clustering algorithms may only suit data with spherically or hyperspherically structures well. For some other certain types of data structures, for instance, linear or hyperplane shaped data clusters, most of the current clustering methods may not perform well [1]. Although some methods like graph-theoretic methods are capable of detecting linear or non-linear cluster structures, there are no explicit representations for the clusters. Thus these algorithms incur difficulties in performing classification tasks. Moreover, in some areas like computer vision and image understanding, clustering algorithms for structure segmentation may not only involve the cluster centers but also need to consider the geometry of the clusters. Last but not least, real data such as the microarray gene expression data often overlap each other. For any clustering algorithms, a challenge how to take both properties of overlapping and the linear subspace structure of the data into account is worth investigating.

It has been known that hyperplanes-based pattern analysis is getting more and more popular and providing researchers with great potential to handle many classification problems [2–5]. One of the most important methods of hyperplanes-based pattern analysis is the approach of support vector machines

(SVMs), which aims to calculate an optimal separating hyperplane between different types of data to obtain the class differentiation [6]. By adopting the quadratic programming and a kernel trick, the SVMs are entitled with excellent capability of pattern classification and have been applied successfully and widely. Despite the success of the SVMs, these algorithms are computationally expensive in many particular applications. Research efforts have recently been made to alleviate the computational burden of the SVMs while maintaining the predictive accuracy by using the hyperplanes-based approximation [7–11]. In these investigations, hyperplanes were used to map each type of data rather than to separate them from each other. The optimal hyperplane minimizes the sum of the squared Euclidean distances of one type data while maximizes the sum of squared Euclidean distances of the other type data. The objective functions are written in the form of the Rayleigh quotient and the solution can be obtained through the generalized eigenvalue decomposition. By using the hyperplanes-based data approximation, the efficiency of the algorithm and the accuracy of classification were reported.

On the other hand, hyperplanes-based clustering techniques are also attracting attentions from research community of pattern recognition[5]. In [2], a k -planes clustering technique which uses hyperplanes to represent cluster centers was presented. The objective of the k -planes clustering is to minimize the sum of the squared Euclidean distances between data points and their projections on their belonging hyperplanes. The k -planes clustering algorithm iteratively updates the partition matrix and clustering hyperplanes until a convergence is reached. The partition matrix is updated by assigning data to its closest hyperplane and the hyperplanes are updated through the eigenvalue decomposition to a covariance matrix. In [3], the authors presented a hyperplane based clustering method which is called the k -bottleneck hyperplane clustering (k -bHPC). The aim of the k -bHPC is to partition data into several groups, and to find a hyperplane for each group that minimizes the maximum distance between the data points to their projections on the hyperplane.

Being motivated by the useful concepts of coupling the fuzzy c -means clustering with hyperplane-based data approximation, a fuzzy mix-prototype clustering technique is presented in this paper where hyperplanes as well as hyperspheres are adopted to approximate cluster centers. The objective function of the proposed clustering is the sum of the distances from all the data samples to the clustering hyperplanes, weighted by the assignment of the point to the corresponding clusters, and penalized by the distances of the data samples to the cluster mass centers. The aim of the mix-prototype clustering is to find a solution to minimize the fuzzy objective function under constraints. The clustering problem can then be viewed as a constrained optimization problem and an iterative algorithm can be obtained by using the Lagrangian multiplier method.

The rest of the paper is organized as the follows. Section 2 describes the proposed fuzzy mix-prototype clustering in detail, including formulation of the fuzzy objective function, derivation of an solution and description of the resulting algorithm. In Section 3, we report the experimental results of the proposed

method and compare these results with those obtained from some existing methods. Concluding remarks of the proposed approach are addressed in Section 4.

2 The Fuzzy Mix-Prototype Clustering

2.1 The FMP Objective Function

Being different from most current clustering techniques such as the c -means and fuzzy c -means, the proposed fuzzy mix-prototype clustering adopts the geometrical hyperplanes $\mathbf{h}_j = (\mathbf{w}_j, v_j)$, $j = 1, \dots, c$ which maximizes the cluster variances, where c is the number of clusters. In the following parts, \mathbf{h}_j will be referred as a hypercluster. To prevent from producing indefinite clusters, a mass center for each hypercluster is also calculated. In the rest parts of the paper, all the vectors are column vectors by default and written in bold. Transpose of a vector or matrix is written with superscript t .

The objective function of the proposed FMP can be written as the follows

$$J_{FMP} = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \left(\gamma \cdot d(\mathbf{x}_i, \mathbf{h}_j) + (1 - \gamma) \cdot d^2(\mathbf{x}_i, \mathbf{g}_j) \right) \quad (1)$$

where $\gamma \in (0, 1)$ is a parameter that penalize the membership of sample points close to hypercluster but far from the mass center, \mathbf{h}_j is the j -th hypercluster (\mathbf{w}_j, v_j) , \mathbf{g}_j is the j -th mass center, and $\mathbf{w}_j = \{w_{1j}, w_{2j}, w_{3j}, \dots, w_{pj}\}$ is a p -dimensional normal vector to the j -th hypercluster. The distance from a data point to the hypercluster is defined as

$$d(\mathbf{x}_i, \mathbf{h}_j) = \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} \quad (2)$$

$$d(\mathbf{x}_i, \mathbf{g}_j) = \|\mathbf{x}_i - \mathbf{g}_j\|_2 \quad (3)$$

$$\|\mathbf{w}_j\| = 1; j = 1, \dots, c; \exists w_{ij} \neq 0 \quad (4)$$

$$\sum_{j=1}^c u_{ij} = 1, i = 1, \dots, n, u_{ij} \in [0, 1] \quad (5)$$

where $\mathbf{w}_j^t \cdot \mathbf{x}_i$ denotes the dot product between vector \mathbf{w}_j^t and vector \mathbf{x}_i .

2.2 An Iterative Solution to FMP

Given the objective function of the FMP to minimize:

$$J_{FMP} = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \left(\gamma \cdot \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} + (1 - \gamma) \cdot \|\mathbf{x}_i - \mathbf{g}_j\|_2^2 \right) \quad (6)$$

which subjects to constraints expressed in Eq. (4) and Eq. (5).

The Lagrangian function of J_{FMP} can be written as

$$L = J_{FMP} - \sum_{j=1}^c \lambda_j (\mathbf{w}_j^t \cdot \mathbf{w}_j - 1) - \sum_{i=1}^n \alpha_i (\sum_{j=1}^c u_{ij} - 1) \quad (7)$$

By taking the first partial derivatives of L with respective to u_{ij} , v_j , w_j and g_j , and setting the equation to 0, the necessary condition for minimizing Eq.(6) can be obtained,

$$u_{ij}^* = \frac{\left(\frac{1}{\gamma \cdot d(\mathbf{x}_i, \mathbf{h}_j) + (1-\gamma) \cdot d^2(\mathbf{x}_i, \mathbf{g}_j)} \right)^{\frac{1}{m-1}}}{\sum_{j=1}^c \left(\frac{1}{\gamma \cdot d(\mathbf{x}_i, \mathbf{h}_j) + (1-\gamma) \cdot d^2(\mathbf{x}_i, \mathbf{g}_j)} \right)^{\frac{1}{m-1}}} \quad (8)$$

$$v_j^* = \frac{\mathbf{w}_j^t \mathbf{X} \mathbf{u}_j^m}{\mathbf{e}^t \mathbf{u}_j^m} \quad (9)$$

$$\mathbf{M}_j = \sum_{i=1}^n u_{ij}^m \mathbf{x}_i (\mathbf{x}_i^t - \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i^t}{\sum_{i=1}^n u_{ij}^m}) \quad (10)$$

$$\mathbf{g}_j^* = \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^n u_{ij}^m} \quad (11)$$

where \mathbf{e} is a n -dimensional column vector with all of its elements equal to one, \mathbf{u}_j^m is the m -th power to \mathbf{u}_j and \mathbf{X} is the p by n data matrix and \mathbf{w}_j is the eigenvector corresponding to eigenvalue $\frac{\lambda_j}{\gamma}$ of matrix \mathbf{M}_j .

2.3 The FMP Algorithm

The computational procedure of the FMP algorithm is an iterative update between \mathbf{U} , \mathbf{h} and \mathbf{g} which is analogous to that of the FCM and can be summarized as follows.

1. Initialize parameters, including the fuzziness parameter m , cluster number c , penalize parameter γ ; set iteration count $k=0$, ε to be a positive small number.
2. Initialize partition matrix $\mathbf{U}(k)$, hyperclusters $\mathbf{h}_j(k)$, $j = 1, \dots, c$ and mass centers $\mathbf{g}_j(k)$.
3. Update $\mathbf{w}_j(k+1)$ through decomposing matrix \mathbf{M}_j in Eq. (10) and selecting the eigenvector corresponding to the smallest eigenvalue.
4. Update $v_j(k+1)$ according to Eq. (9) and $\mathbf{h}_j(k+1)$.
5. Update $\mathbf{g}_j(k+1)$ according to Eq. (11).
6. Update the fuzzy partition matrix $\mathbf{U}(k+1)$ according to Eq. (8).
7. If the algorithm the maximum change in the partition matrix between two successive iterations is less than ε , then terminate the iteration. Otherwise, go to Step 3.

3 Experiments

To validate the proposed FMP clustering, we conducted a group of experiments on MLL-Leukemia data set. The results obtained from the FMP were compared with those obtained from some related methods including the FCM and kernel FCM with Gaussian kernel.

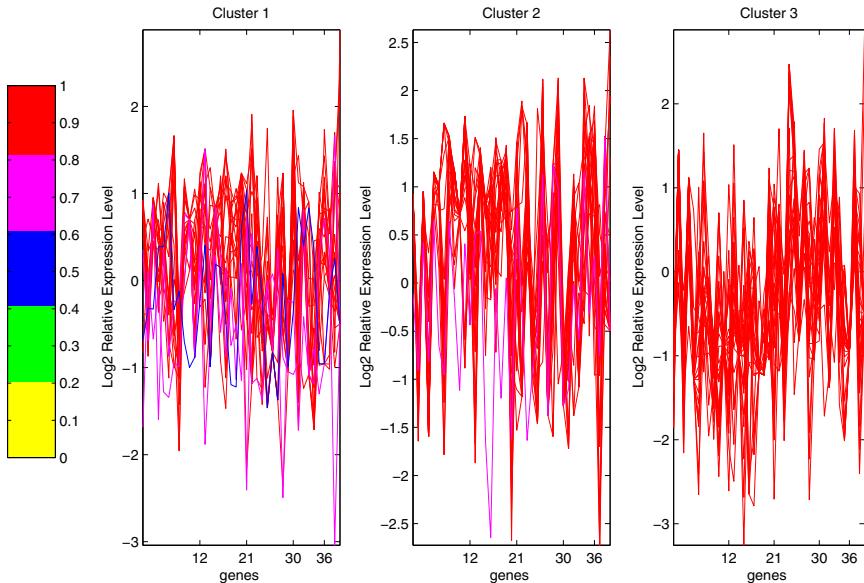


Fig. 1. Gene expression trajectory of FCM clustering on Leukemia data set

The MLL-Leukemia data set consists of 72 samples with 12582 gene expressions in three types of leukemia: 24 acute lymphoblastic leukemia (ALL), 20 mixed-lineage leukemia (MLL) and 28 acute myelogenous leukemia (AML) [12]. The original data set is of high dimensions. For the convenience of data analysis, feature selections and dimension reduction were adopted to remove redundant and irrelevant features before validations were carried out. After these processing steps, the final data set rendered for analysis consisting of 72 samples with 39 genes.

Figures 1-3 showed the clustering results produced by FCM, KFCM and FMP algorithms, respectively. From these Figures, it could be seen samples were partitioned to groups in which they produced the largest membership values. The expression trajectories in each group are plotted in different colors, corresponding to the different ranges of the membership, where the color was red if the membership of the gene to the cluster was in the range of $(0.8, 1]$, purple if the

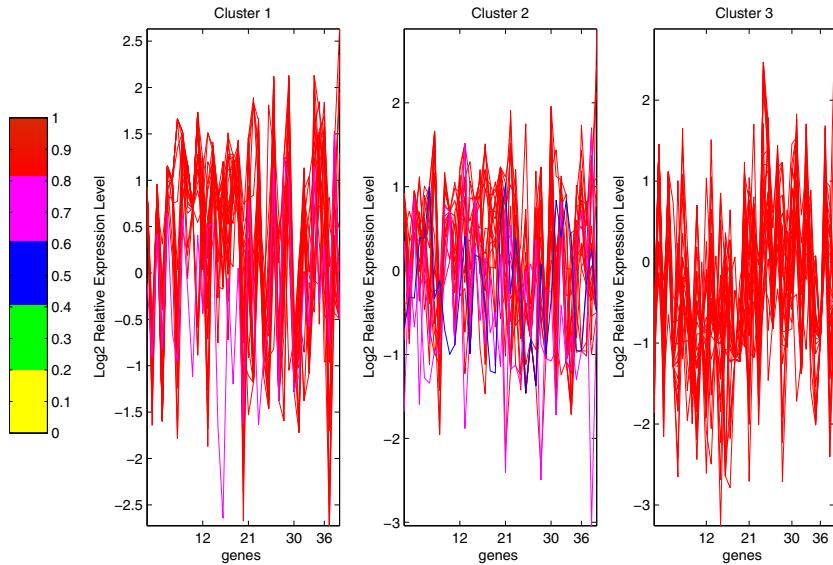


Fig. 2. Gene expression trajectory of KFCM clustering with Gaussian kernel on Leukemia data set

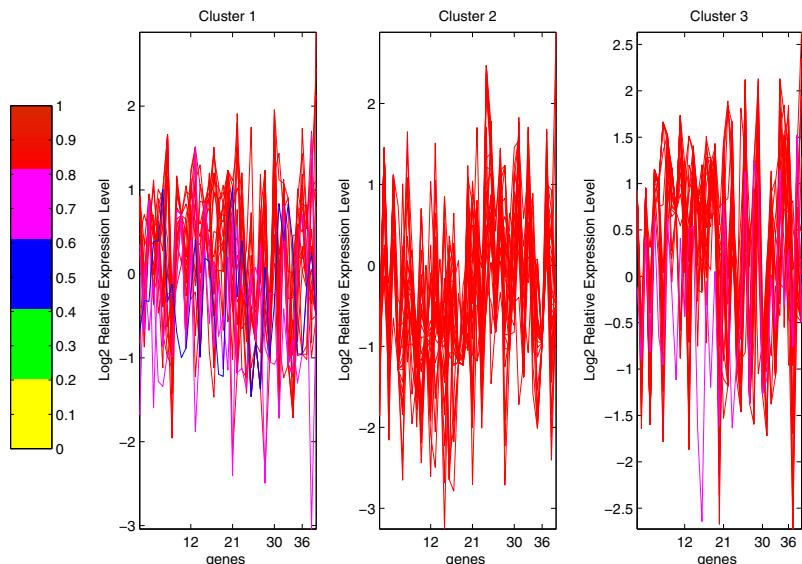


Fig. 3. Gene expression trajectory of FMP clustering on Leukemia data set

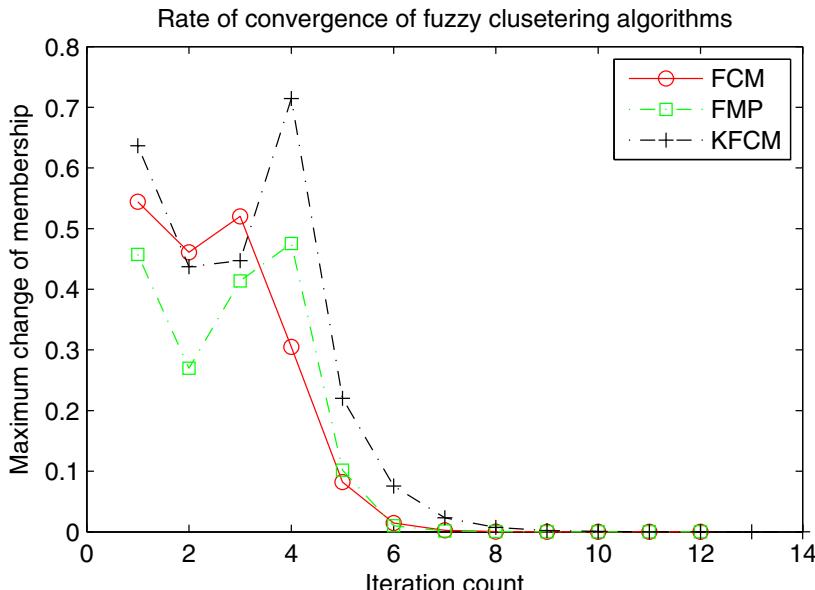


Fig. 4. Rate of convergence of different clustering algorithms on MLL-Leukemia data set

membership was in the range of $(0.6, 0.8]$, blue if the membership was in the range of $(0.4, 0.6]$, green if the membership was in the range of $(0.2, 0.4]$ and yellow if the membership was in the range of $(0, 0.2]$.

Convergence of different clustering algorithms on Leukemia data set were also studied. The results were shown in Figure 4, from which we can see that the values of the 3 algorithms dived sharply at the first 2 iterations, and then climbed to local high after 3 or 4 iterations. The values of the 3 algorithms then declined with small fluctuation and after about 7 iterations, FCM and FMP reached their convergence, followed by KFCM at about 9 iterations.

4 Conclusion

We have presented a proposed fuzzy mix-prototype clustering algorithm, which can be useful for pattern classification in microarray data analysis. We formulated the objective function for the proposed FMP clustering and derived an iterative numerical solution for minimizing the objective function under the given constraints. The proposed method was then applied to analyze Leukemia microarray gene expression data set, and the results were compared against related existing clustering methods including the FCM and KFCM algorithms. The experimental results have demonstrated that the proposed method is an effective alternative for microarray data analysis.

Acknowledgments. The paper is supported by the National Natural Science Foundation of China (Grant No. 61303182), the Natural Science Foundation of Jiangsu Province (Grant No. BK20130210), Fundamental Research Funds for the Central Universities (Grant No. 2012QN17), the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20120095120026), the Postdoctoral Science Foundation of China (Grant No. 2012M521144), the Postdoctoral Science Foundation of Jiangsu Province (Grant No. 1301120C).

References

1. Bezdek, J.C., Coray, C., Gunderson, R., Watson, J.: Detection and characterization of cluster substructure i. linear structure: fuzzy c -lines. *SIAM Journal on Applied Mathematics* 40(2), 339–357 (1981)
2. Bradley, P.S., Mangasarian, O.L.: k -plane clustering. *J. Global Optimization* 16(1), 23–32 (2000)
3. Dhyani, K., Liberti, L.: Mathematical programming formulations for the bottleneck hyperplane clustering problem. In: Le Thi, H.A., Bouvry, P., Pham Dinh, T. (eds.) *Modelling, Computation and Optimization in Information Systems and Management Sciences*, vol. 14, pp. 87–96. Springer, Heidelberg (2008)
4. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge (2000)
5. Liu, J., Pham, T.D.: Fuzzy Hyper-Prototype Clustering. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2010, Part I. LNCS*, vol. 6276, pp. 379–389. Springer, Heidelberg (2010)
6. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2), 121–167 (1998)
7. Fung, G., Mangasarian, O.L.: Proximal support vector machine classifiers. In: ACM SIGKDD KDD 2001: Proc. of the Seventh International Conference on Knowledge Discovery and Data Mining, pp. 77–86. ACM, New York (2001)
8. Fung, G.M., Mangasarian, O.L.: Multicategory proximal support vector machine classifiers. *Mach. Learn.* 59(1-2), 77–97 (2005)
9. Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(1), 69–74 (2006)
10. Yang, X., Chen, S., Chen, B., Pan, Z.: Proximal support vector machine using local information. *Neurocomputing* 73(1-3), 357–365 (2009)
11. Ghorai, S., Mukherjee, A., Dutta, P.K.: Nonparallel plane proximal classifier. *Signal Processing* 89(4), 510–522 (2009)
12. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature* 30(1), 41–47 (2002)

Blind Image Quality Assessment with Semi-supervised Learning and Fuzzy Logic

Ning Mei, Fei Gao, Wen Lu, and Xinbo Gao

School of Electronic Engineering, Xidian University, Xi'an, 710071, China

{ningmei1989,gaofeihifly,luwen.xidian}@gmail.com,
xbgao@mail.xidian.edu.cn

Abstract. Blind image quality assessment (BIQA) is a challenging task due to the difficulties in extracting quality-aware features and modeling the relationship between the features and the visual quality. In this paper, we propose a semi-supervised and fuzzy (S^2F) framework for BIQA. First, we formulate the fuzzy process of subjective quality assessment by using fuzzy logic. Secondly, we introduce the semi-supervised local linear embedding (SS-LLE) to learn the mapping from the features to the truth values using both the labeled and unlabeled images. Experimental results on two benchmarking databases demonstrate the effectiveness and promising performance of the proposed S^2F framework for BIQA.

Keywords: Blind image quality assessment, Fuzzy logic, Semi-supervised LLE, Natural scene statistics.

1 Introduction

Blind image quality assessment (BIQA) has attracted increasing attentions during the past decades. Due to the limited exploration of human visual system (HVS), it is a challenging task to model the relationship between the image features and the visual quality. It is therefore of great difficulty to develop effective BIQA metrics, especially universal BIQA (UBIQA) metrics.

The past five years have witnessed the emergence of various new BIQA algorithms [1–7]. Given natural scene statistics (NSS) features, BLIINDS-II [1] relies on a simple Bayesian inference model to predict image quality, and DIIVINE [2] bases on a two-stage framework which adopts distortion-specific BIQA regression metrics to estimate the image quality. BRISQUE [3] introduced a new spatial NSS features to quantify possible losses of “naturalness” in the image.

Yet there are two drawbacks of these algorithms. First, only the labeled images are adopted for training. However, unlabeled data can improve the learning performance [11]. In addition, these metrics try to learn a direct mapping from the features to the quality. However, human perception would rather be a fuzzy process than a discriminative one. To overcome these problems, we propose a semi-supervised and fuzzy (S^2F) framework for BIQA in this paper. In the proposed framework, we first formulate the fuzzy process of subjective quality assessment; and then introduce the semi-supervised local linear embedding

(SS-LLE) to learn the mapping from the NSS features to the truth values using both the labeled and unlabeled images. Experimental results demonstrate the effectiveness performance of the proposed S²F framework for BIQA.

The rest of this paper is organized as follows. Section II details the proposed S²F metric. Section III presents the experiments conducted on the LIVE database II and the TID2008 database. Finally, Section IV concludes this paper.

2 Semi-supervised and Fuzzy Framework for Blind Image Quality Assessment

Fig. 1 shows the proposed semi-supervised and fuzzy framework. Firstly, we extract image features based on NSS. Secondly, we formulate the fuzzy process of subjective quality assessment and convert the quality score to several truth values. Finally, we introduce SS-LLE to learn the mapping from features to truth values, and the quality scores of the test images are estimated based on them.

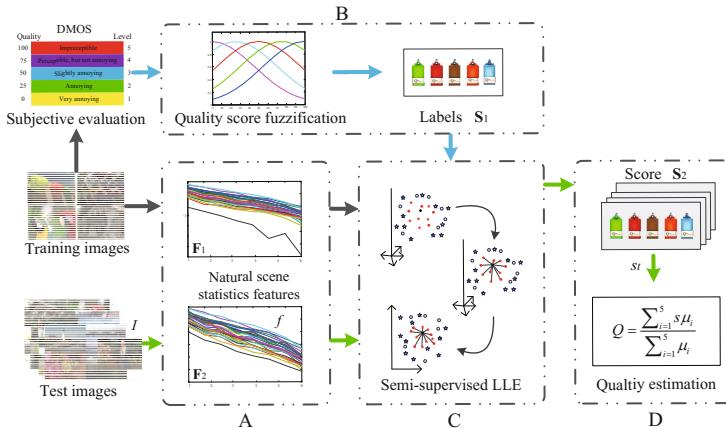


Fig. 1. The proposed semi-supervised and fuzzy logic framework for BIQA

2.1 Natural Scene Statistics Features

In order to capture the statistical properties of images, we utilize the wavelet transform owing to its scale, space and orientation selectivity. An input image I is decomposed into three scales wavelet coefficients. We combine the same subbands through averaging which makes the number of subbands into six. For each subband, we extract the magnitude feature m_k to encode the generalized spectral behavior, and the entropy feature e_k to represent the generalized information:

$$m_k = \frac{1}{N_k \times M_k} \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} \log_2 |C_k(i, j)|, \quad (1)$$

$$e_k = \sum_{j=1}^{N_k} \sum_{i=1}^{M_k} p[C_k(i, j)] \ln p[C_k(i, j)], \quad (2)$$

where M_k and N_k ($k = 1, 2, \dots, 6$) are the length and width of the k -th subband respectively, and $C_k(i, j)$ stands for the (i, j) coefficient of the k -th subband. Stack these 12 statistics to form a single vector

$$f = [m_1, m_2, \dots, m_6, e_1, e_2, \dots, e_6]^T. \quad (3)$$

The relationship between the features and the quality is visualized in Fig. 2. The magnitude spectra of the reference images have the similar exponential decay characteristics across scales while the distorted ones have different downtrend. The distribution of magnitudes follows the law that the increase of distortion degree brings out the sharper decrease [5]. In this paper, we apply this speciality to conduct semi-supervised manifold.

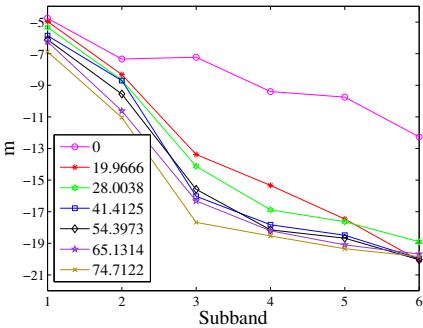


Fig. 2. Magnitudes of the wavelet coefficients for different DMOS decay exponents across six subbands

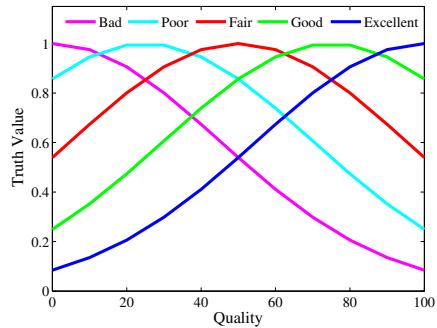


Fig. 3. The proposed fuzzy membership functions for five categories of image qualities

2.2 Quality Score Fuzzification

The term "fuzzy logic" was introduced in fuzzy set theory proposed by Zadeh [13]. The basic principle of fuzzy logic is a matter of degree and it deals with reasoning that the real world is approximate rather than fixed and exact [12].

Since human is the termination of all multimedias, subjective quality assessment is the most reasonable criterion for IQA. The recommendation of subjective quality assessment, ITU-R 500-11 [8], is a five-grade impairment scale: "excellent", "good", "fair", "poor", and "bad". Therefore the fuzzy modeling of image quality is somewhat intermediate, which can reduce the granularity of the image's characterization. Meanwhile, Gaussian function is commonly used to model the fuzzy membership. As a result, we design the Gaussian-based fuzzy membership functions of which the mean value represents primary term and the variance details the possible distribution. The fuzzy membership function acts on difference mean opinion scores (DMOS), shown in Fig. 3.

$$s_l = f(DMOS) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(DMOS-\mu_l)^2}{2\sigma^2}} (l = 1, 2, 3, \dots, 5), \text{ and}$$

$$s = [s_1, s_2, \dots, s_5]^T. \quad (4)$$

The fuzzy expression models the process of human perception. The characterization of the image is better in five truth values than in one score which makes the granularity of image smaller and the volumes of information larger.

2.3 Semi-surervised Local Linear Embedding

In BIQA problem, we cannot grade an image directly but can easily sequence it in a large amount of images. The clue to this phenomenon is that perception is manifold. Therefore, we inherit the advantages of SS-LLE to BIQA which is capable and competent in formulating the process of human perception and learning the mapping from features to the truth values.

The SS-LLE consists of three steps shown in Table 1.

Table 1. Algorithm of SS-LLE

Input: Feature vector f_i and truth value s_i

Stack all the truth values to form \mathbf{S} . N is the total number of images.

$$\mathbf{S} = [s_1, s_2, \dots, s_N].$$

Step 1 Find k nearest neighbors for each feature points f_i .

Step 2: Compute the reconstruction coefficients

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^N \left\| \mathbf{f}_i - \sum_{j=1}^k \mathbf{W}_{ij} \mathbf{f}_j \right\|^2,$$

Subject to constraint: $\sum_{j=1}^N W_{ij} = 1 (i = 1, 2, \dots, N)$.

Step 3: Compute the low dimensional embedding.

$$\phi(\mathbf{S}) = \sum_{i=1}^N \left\| \gamma_i - \sum_{j=1}^k \mathbf{W}_{ij} \gamma_j \right\|_2^2 = \mathbf{S} \mathbf{M} \mathbf{S}^T,$$

Subject to two constraints: $\sum_{i=1}^N \gamma_i = 0$ and $\frac{1}{N} \sum_{i=1}^N \gamma_i^T \gamma_i = \mathbf{I}$,

$\mathbf{M} = (\mathbf{1} - \mathbf{W})^T (\mathbf{1} - \mathbf{W})$, \mathbf{M} is partitioned into four parts, and \mathbf{S} is partitioned into two parts, referred to labeled and unlabeled images, and find the smallest $d + 1$ eigenvectors of matrix \mathbf{M} :

$$\min_{\mathbf{S}_2} [\mathbf{S}_1, \mathbf{S}_2] \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{12}^T & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1^T \\ \mathbf{S}_2^T \end{bmatrix},$$

set the gradient of the above objective function to zero

$$\mathbf{M}_{22} \mathbf{S}_2^T = \mathbf{M}_{12} \mathbf{S}_1^T, \text{ and}$$

$$\mathbf{S}_2 = \left(\mathbf{M}_{12} \mathbf{S}_1^T / \mathbf{M}_{22} \right)^T$$

Output: Unlabeled truth values \mathbf{S}_2

As a part of unlabeled images, the test image participates in SS-LLE. \mathbf{S}_2 contains the five truth values associated with the test image. It is reasonable

to contain some test images in the training stage when the labeled images are insufficient, and this kind of semi-supervised approach will make the quality assessment more mature with the training repeated.

2.4 Quality Estimation

In this section, we introduce two different approaches for quality estimation. The first one includes a defuzzification module to keep in step with traditional IQA metrics. However, practical applications call for non-numeric descriptors rather than one quality score. By comparison, the second approach takes the five truth values obtained from SS-LLE for the final result. For clarity, the first approach is referred as S²F-I and the second is referred as S²F-II.

S²F-I: To obtain the traditional quality of images, we should defuzzify the five truth values obtained in learning module. We adopt the center of area (CoA) defuzzification method to interpret the membership degrees into a real value. The formula of CoA is

$$Q = \text{defuzz}(s_t) = \frac{\sum_{i=1}^5 s_t \mu_i}{\sum_{i=1}^5 \mu_i}. \quad (5)$$

As stated earlier, s_t are the output labels of SS-LLE for the test image. The outcome Q , ranging from 0 to 100, is the final quality score of the test image.

S²F-II: The subconscious feeling of an image the moment we see it, is about good feasibility which is expressed in words rather than exact score. In view of the practical application, we directly adopt the estimated five truth values as the presentation of the image quality.

$$Q = s_t. \quad (6)$$

The five truth values for each of the five functions represent the degrees of truth they belongs to “excellent”, “good”, “fair”, “poor”, and “bad”. This estimation method makes the assessment much closer to human behavior.

3 Experiments and Analysis

To verify the effectiveness of the proposed S²F-I and S²F-II metric, we test them on two benchmarking databases: the LIVE database II [9] and the TID2008 database [10]. The LIVE database II consists of 29 reference images and 779 distorted images. The TID2008 database contains 1700 test images and 25 reference images over 17 distortion categories.

The criteria considered in experiment are the Spearman's rank ordered correlation coefficient (SROCC) and the linear correlation coefficient (LCC). A value close to 1 indicates superior correlation with human perception. In original LLC calculation for S²F-I, each point in the group of quality represent a test image. However, in the LCC calculation for S²F-II, five points cooperate in presenting an image. This makes the LLC changes from a 1-to-1 calculation to 5-to-5 calculation. Several experiments is detailed in the following subsections to verify the consistency between the proposed BIQA metrics and subjective IQA.

3.1 Consistency Experiments

We randomly select 23 groups of the LIVE database II for labeled images and the rest for unlabeled, and repeat the training 1000 times to evaluate the average performance. We compare the proposed metrics with not only BIQA metrics, i.e. BLIINDS-II [1], DIIVINE [2], BRISQU [3], NIQE [4], SRNSS [5], and SF100 [6], but also two full-reference (FR) IQA metrics, PSNR and SSIM [7]. The number of nearest neighbors k selected in SS-LLE is 70. The scale of fuzzy membership function l is 5. The variance of Gaussian distribution σ is 90.

Because of the difference in the indices calculation, it is unreasonable to compare S^2F -II with other available BIQA methods, but feasible with S^2F . In this subsection, we present the performances of S^2F -I and S^2F -II separately.

The Performance of S^2F -I. Figure 4 shows the scatter plots and the nonlinear curve fittings between the estimated quality scores by S^2F -I and DMOS across the entire test set. Table 2 shows the LCC and SROCC of both S^2F -I and S^2F -II. It can be seen that S^2F -I achieves the highest LCC and SROCC, outperforming the other methods which demonstrates the effectiveness of S^2F -I. The process of training and test (without feature extraction) on the entire LIVE using S^2F -I takes around 21.66 s with a 2.66 Ghz PC with 2 GB of RAM.

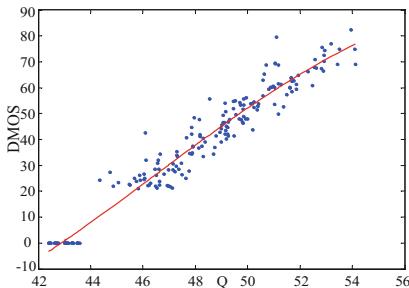


Fig. 4. Predicted Q vs. subjective DMOS on the entire LIVE database II

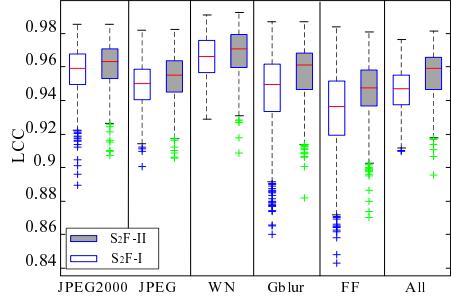


Fig. 5. Boxplot of LCC of S^2F -I and S^2F -II in 5 scales on LIVE

The Performance of S^2F -II. The comparation of S^2F -I and S^2F -II presented in Table 2 demonstrates the advancement of S^2F -II. It is notable from Fig. 5 that the performance of S^2F -II is much better than S^2F -I for every distortion. The reason behind this improvement is that truth values reduce the granularity of an image and better imitate the human visual perception.

3.2 Robustness to Parameters

To verify how the parameters affect the performance of S^2F , we conduct experimental studies on four influential parameters: the form of membership function, l , σ and k . The experiment results show that the proposed frameworks keep performing well for different selection of parameters. Therefore, the parameters in our framework are insensitive which can be decided in accordance with actual condition of applications.

Table 2. Median LCC and SROCC of different metrics on the LIVE database II

Metric	JP2K		JPEG		WN	
	LCC	SROCC	LCC	SROCC	LCC	SROCC
PSNR	0.8962	0.8726	0.9858	0.8839	0.8895	0.9415
SSIM	0.9367	0.9389	0.9695	0.9283	0.9428	0.9635
DIIVINE	0.9220	0.9130	0.9880	0.9100	0.8880	0.9840
BLLIINDS-II	0.9630	0.9293	0.9854	0.9421	0.9436	0.9693
SRNSS	0.9359	0.9283	0.9404	0.9306	0.9473	0.9382
BRISQUR	0.9229	0.9139	0.9851	0.9647	0.9093	0.9786
NIQE	0.9370	0.9172	0.9773	0.9382	0.9128	0.9662
SF100	0.9301	0.9248	0.9782	0.9310	0.8880	0.9622
S²F-I	0.9578	0.9354	0.9668	0.9183	0.9370	0.9710
S²F-II	0.9616	0.9355	0.9693	0.9296	0.9474	0.9604
Metric	Gblur		FF		Entire database	
	LCC	SROCC	LCC	SROCC	LCC	SROCC
PSNR	0.7834	0.8839	0.8895	0.7646	0.8240	0.8636
SSIM	0.8740	0.9283	0.9428	0.8942	0.8634	0.8834
DIIVINE	0.9230	0.9100	0.8880	0.9210	0.9170	0.9170
BLLIINDS-II	0.9481	0.9421	0.9436	0.9232	0.9232	0.9202
SRNSS	0.9356	0.9306	0.9473	0.9327	0.9318	0.9304
BRISQUR	0.9506	0.9647	0.9093	0.9511	0.9424	0.9395
NIQE	0.9525	0.9382	0.9128	0.9341	0.9147	0.9135
SF100	0.9516	0.9310	0.8880	0.9614	0.9213	0.9214
S²F-I	0.9556	0.9183	0.9370	0.9367	0.9464	0.9412
S²F-II	0.9572	0.9296	0.9474	0.9365	0.9560	0.9360

3.3 Robustness to Database

In order to demonstrate the algorithm is database independence, we train S²F-I on the LIVE database II and test only on the four distortions that it is trained for: JPEG, JP2K, WN and Gblur of TID2008 as done in [1–3]. Table 3 shows the comparison result on TID2008 . The SROCC of S²F-I metric drops in other because of the differences in simulated distortions and objective evaluation which makes the fuzzy module not precise. However, the performance of S²F-I is still encouraging compared with the available methods. Therefore, the proposed S²F-I is robust against the test data and can be applied to other different databases.

Table 3. Median SROCC OF different metrics trained on the LIVE database II and tested on TID2008 database

	JP2K	JPEG	WN	Gblur	All
PSNR	0.8250	0.8760	0.9230	0.9342	0.8700
SSIM	0.9603	0.9354	0.8168	0.9544	0.9016
BLLIINDS-II	0.9157	0.901	0.6600	0.8500	0.8442
DIIVINE	0.924	0.966	0.851	0.862	0.889
BRISQUR	0.832	0.924	0.82	0.881	0.896
S²F-I	0.9115	0.9143	0.7503	0.8422	0.8761

4 Conclusions

In this paper, we propose a new semi-supervised and fuzzy framework for BIQA, called S²F. Experimental results on two benchmarking databases demonstrate that S²F makes an obvious improvement over state-of-the-art BIQA metrics. Nevertheless, the proposed framework is still limited while compared with the best FR-IQA metrics. Improvement and development will be our future work.

Acknowledgements. This research was supported partially by the National Natural Science Foundation of China (Grant Nos. 61125204, 61001203 and 61172146), the Fundamental Research Funds for the Central Universities (Grant No. K5051202048), and Shaanxi Innovative Research Team for Key Science and Technology (No.2012KCT-02).

References

1. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.* 21, 3339–3352 (2012)
2. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.* 20, 3350–3364 (2011)
3. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 4695–4708 (2012)
4. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.* 22, 209–212 (2013)
5. He, L., Tao, D., Li, X., Gao, X.: Sparse representation for blind image quality assessment. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1146–1153 (2012)
6. Ye, P., Kumar, J., Kang, L., Doermann, D.: Real-time no-reference image quality assessment based on filter learning. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (to be published, 2013)
7. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612 (2004)
8. ITU-T Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland (2004)
9. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Live image quality assessment database release 2, <http://live.ece.utexas.edu/research/quality>
10. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: Tid2008 - a database for evaluation of full-reference visual quality assessment metrics. *Adv. of Modern Radioelectron.* 10, 30–45 (2009)
11. Yang, X., Fu, H., Zha, H., Barlow, J.L.: Semi-supervised nonlinear dimensionality reduction. In: Proceedings of International Conference on Machine Learning, pp. 1065–1072 (2006)
12. Zadeh, L.A.: Fuzzy logic. *IEEE Comput. Mag.* 21, 83–93 (1988)
13. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)

Cross-View Action Recognition via Bilingual Bag of Dynamical Systems

Changhong Chen^{*}, Shunqing Yang, and Zongliang Gan

Nanjing University of Posts and Telecommunications, Nanjing, China
chenchh@njupt.edu.cn

Abstract. In this paper, a new framework is proposed for cross-view action recognition. Spatio-temporal patches are extracted as low-level feature and each patch is represented as a linear dynamical system (LDS). Bag of dynamical systems (BODS) is employed for middle-level representation. In order to bridge different views, we transform BODS pairs into a bilingual BODS through transferable dictionary pairs. Bilingual dictionaries are learned for the source and target view, which guarantee that the same action from the two views have same high-level representation. Support vector machine (SVM) is employed as the classifier. The experimental results on the IXMAS multi-view dataset show the effectiveness of proposed algorithm compared with others. The performance on the top view is also excellent.

Keywords: Cross-view action recognition, transfer learning, bilingual bag of dynamical systems.

1 Introduction

Human action recognition aims to recognize the actions of one or more agents from a series of observations. Appearance-based methods are widely applied to action recognition. It is very important to extract visual representations for the actions, such as shape features [1, 2], space-time templates [3, 4] and motion flow patterns [5]. These features perform well in recognizing actions with limited view variations, but tend to be powerless for large view variations. The major reason lies in the obvious changes in the appearance of actions with different viewpoints. It is even difficult for people to recognize them when the viewpoints change greatly, such as the horizontal view and top view. As a result, appearance-based methods using low-level features become less discriminative.

An interesting algorithm is proposed by Junejo et al. [6], which relies on weak geometric properties. An simple and interesting action representation called self-similarity descriptors is presented in this paper. It is found to be highly stable under view changes. This method is satisfying in most cases. However, the experimental results show the approach performs poorly when the top view as source or target view.

* Corresponding author.

Another encouraging work was also proposed by Farhadi et al. [7]. They train a discriminative aspect model, called latent bilinear model, to handle the view invariance. This model is effective for objects and human activities in views, but it requires good parameter initialization, which is not easy to be obtained.

Transfer learning is more popular and effective than the aforementioned two algorithms. Given a pair of views, the features are learned from them and construct a bridge between them. For a new action class observed in one view, transfer learning enables recognition in the second view through the bridge. Farhadi and Tabrizi [8] employed Maximum Margin Clustering to generate split-based features in the source view and a classifier is trained to predict split-based features in the target view. Liu et al. [9] extract bilingual-words from a pair of views through co-cluster two vocabularies (high level features) into visual-word clusters. The source and target views are explicitly connected by a smooth virtual path represented as a sequence of linear transformations of action descriptors in [10]. Li et al. [11] introduce Hankelets as a new feature and learning bilingual Hankelets through transfer learning. The most encouraging method is proposed by Zheng et al. [12], who learn the two dictionaries of source and target views simultaneously to ensure the same action to have the same representation. Although these transfer learning algorithms achieve good performance. However, they are harder to transfer action models across views that involves the top view.

Transfer learning is always combined with bag of words (BOW)[13], which is effective in constructing codebook for video sequence. The codebook is constructed by the k-means clustering algorithm, which is based on the similarity of the detected spatio-temporal patches. Such method does not perform well under different viewpoint because of the great changes in the appearance. The bag of dynamical system (BODS) method proposed by Ravichandran et al. [14] is a good choice to model the spatio-temporal patches. This BODS representation is analogous to the BOW, except that linear dynamical systems (LDSs) is used as feature descriptors. This method shows its effectiveness in recognizing dynamic textures in challenging scenarios.

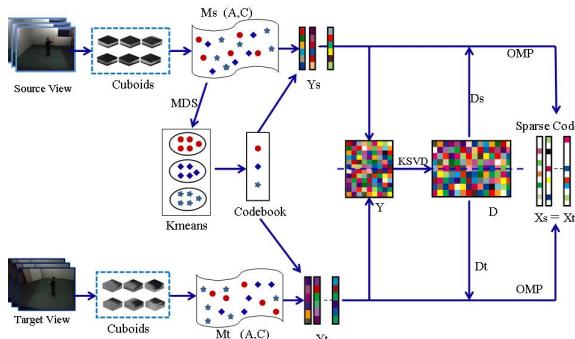


Fig. 1. The framework of the proposed bilingual BODS

Motivated by the success of transfer learning and BODS, we present a new algorithm in this paper, called bilingual BODS. The framework of our work is shown in Fig.1. Spatio-temporal patches are extracted and modeled as LDSs. The codebook are then formed based on clustering the LDSs on non-Euclidean manifold. In order to building the bridge under two different views, bilingual dictionaries are learned simultaneously under the same guideline as [12]. As a result, the same bilingual BODS representations can be obtained for the same action under different views. The proposed bilingual BODS build the bridge for different views and obtained wonderful experiment results on the IXMAS multi-view dataset [15].

2 Low-Level Feature Extraction

Many strong spatio-temporal feature detection methods has been proposed, such as Harris3D detector [16], cuboids detector[17], Hessian detector [18], dense sampling [19]. In this paper, we combine cuboids detector [17] and dense sampling [19] for feature extraction. Cuboids detector are employed to extract the action information [17]. Interest points of the video are detected by separable linear filters. 2D Gaussian filter is applied along the spatial dimensions and 1D Gabor filter applied temporally.

Dense features obtained regular sampling of local spatio-temporal patches has been proved to be more efficient for human action recognition [20]. Besides the cuboids detector, we also extract dense 3D patches at regular positions with the same scale as the cuboid detector.

3 Bilingual Bag of Dynamical Systems

3.1 Linear Dynamical System

After extracting the spatio-temporal feature $\{F_t\}$, a LDS is built as:

$$\begin{cases} x_t = Ax_{t-1} + v_t, \\ F_t = Cx_t + w_t, \end{cases} \quad (1)$$

where x_t is the hidden state at time t , A is the state-transition matrix and C is the observation matrix. The state and observation noises submit to Gaussian distribution, which are given by $v_t \sim N(0, Q)$ and $w_t \sim N(0, R)$, respectively. The choice of the parameters is not unique and a closed-form solution was also proposed in [21]. The closed-form solution commonly uses principal component analysis (PCA) to get the observation function. The columns of C are the principal components of the image sequence and the state vector x_t is a set of PCA coefficients.

The parameters of spatio-temporal features $M = (A, C)$ are used as the descriptors.

3.2 Bag of Dynamical Systems

Traditional cluster algorithms are not suitable for the learned LDSs directly, because the LDSs parameters are lie on a non-Euclidean manifold. Non-Linear Dimensionality Reduction (NLDR) should be applied to find a low-dimensional Euclidean embedding of the descriptors (A, C) .

Given the descriptors set $\{M_i\}_{i=1}^T$, where T is the number of features extracted from the video sequences. The distances between the descriptors $M_1 = \{C_1, A_1\}$ and $M_2 = \{C_2, A_2\}$ are calculated through Martin distance [22].

$$d^2(M_1, M_2) = -\log \prod_{i=1}^n \cos^2 \theta_i, \quad (2)$$

where θ_i is the i^{th} principle angle between the observation matrices. Once the pairwise distances are available, Multidimensional Scaling (MDS) can be used to obtain low-dimensional embedding $\{e_i\}_{i=1}^T$, where k-means cluster algorithms can be applied to build the codebook.

Suppose K cluster centers are $\{K_i\}_{i=1}^K$. The centers are not correspond to the descriptors (A, C) . In order to obtain the codewords $\{W_i\}_{i=1}^K$, we choose the corresponding systems whose low-dimensional representation is closest to the cluster center as [14].

$$W_i = M_p, \quad p = \arg \min_j \|M_j^e - e_i\|^2. \quad (3)$$

The dynamical system M can be connected with the codebook according to:

$$k = \arg \min_i d_M(M, W_i), \quad i \in \{1, \dots, K\} \quad (4)$$

3.3 Middle-Level Representation Based on Soft Weighting

Each video sequence can be represented using this codebook as a histogram: $y = [y_1, y_2, \dots, y_K]^T$. The simplest representation is obtained by the average occur times of the codewords. Suppose the codeword W_m occurs C_m times in the video, y_m can be represented as:

$$y_m = \frac{C_m}{\sum_{m=1}^K C_m}, \quad m = 1, \dots, K. \quad (5)$$

This approach is the term frequency weighting. It treats each codeword independently and overlooks the inherent linguistics among the words. Furthermore, it cannot capture the inter-word relationship. For instance, two spatio-temporal patches are both assigned to a codeword. However, they are not equally similar to the codeword, which can not be reflected by the term frequency.

In this paper, we borrow the idea of soft-weighting [23]. It contains two key steps: the assignment of keypoint-to-word is one-to-many and the importance of the codeword is decided by their linguistic relationship. For the codebook W , each dynamical system is assigned to k nearest codewords instead of one nearest codeword. The soft-weight of a word W_m in a action sequence, denoted as y_m , is measured as:

$$y_m = \sum_{i=1}^k \sum_{j=1}^{L_i} p(i) \cdot \text{sim}(j, m), \quad (6)$$

where L_i is the set of dynamical systems whose i^{th} nearest neighbor is W_m , $\text{sim}(j, m)$ represents the similarity between the dynamical system M_j and the codeword W_m . The $p(i)$ is a function to quantify the importance of $\text{sim}(j, m)$, which simply represent $p(i) = \frac{1}{2^{i-1}}$ in this paper. That is, it contains the same value for different dynamical system M_j in the same video sequence.

3.4 Bilingual Codebook Formulation

We employ the transformable dictionary pair representation[12] for the bilingual codebook formulation. Suppose we have N video sequences shared in the source and target views. Y_s and Y_t are the middle-level representations of them. Our goal is to find a sparse representation X , which can denotes Y_s and Y_t simultaneously. It can be realized by designing bilingual dictionaries D_s and D_t for the source and target views, which can be realized by:

$$\arg \min_{D_s, D_t, X} \|Y_s - D_s X\|_2^2 + \|Y_t - D_t X\|_2^2 \quad s.t. \forall i \quad \|x_i\|_0 \leq s, \quad (7)$$

Where the first term is the reconstruction error of the source view and the second term is that of the target view. x_i is the i^{th} representation of X and $\|x_i\|_0 \leq s$ is the sparse constraint. The bilingual dictionaries $\{D_s, D_t\}$ can be learned using the K-SVD algorithm[24], as shown in Fig.1.

Given the bilingual dictionaries, the high-level sparse representation X can be solved by orthogonal matching pursuit (OMP) algorithm [25].

4 Numerical Experiments

In this section we evaluate our algorithm with the IXMAS multi-view action dataset [15], which contains 11 daily-live actions performed each 3 times by 12 actors taken from 5 different views: four side views and one top view.

Spatio-temporal patches are extracted by cuboids detector [17] and dense sampling [19]. The cuboids number is not fixed and is larger for intense actions. All the cuboids are modeled by LDSs and represented by the BODS. After learning the bilingual codebooks, the same action from both the source and the target views can obtain similar high-level sparse representation.

Because there are only 11 actions in the database, we adapt leave-one-out for the experiments. For each view pair, 10 actions are used for learning the bilingual codebooks. The high-level sparse representation of all actions can be obtained by the codebooks in separate viewpoint. The representation of the left action in the target view can be compared with these representations of all actions in the source view by support vector machine (SVM).

The experimental results are shown in Table 1. Our experimental results is encouraging. All the recognition results are over 97%. Although top view samples have great difference with horizontal view samples, our algorithm have no error for the top view as the target view. We also compared our average recognition results with other approaches as shown in Fig.3. The approach of [6] is based on self-similarity descriptors and don't obtain satisfying performance compared with other transfer learning algorithms. The recognition results of [12] and the proposed algorithm are better than the other two algorithms [9,11]. For Cam0, Cam2 and Cam3 as the target views, our algorithm has comparable excellent results as [12]. For other two target views, our algorithm is obviously better than [12]. When the top view Cam4 as the target view, the proposed bilingual BODS achieve 100% accuracy. The excellent performance testify the validity of the proposed bilingual BODS for cross-view recognition.

Table 1. Performance of the proposed bilingual BODS

	Target View					
	(%)	Cam0	Cam1	Cam2	Cam3	Cam4
Source View	Cam0		99.49	99.75	99.75	100
	Cam1	100		100	100	100
	Cam2	98.99	100		99.75	100
	Cam3	100	100	100		100
	Cam4	99.49	100	97.73	99.49	
	Ave.	99.62	99.87	99.37	99.74	100

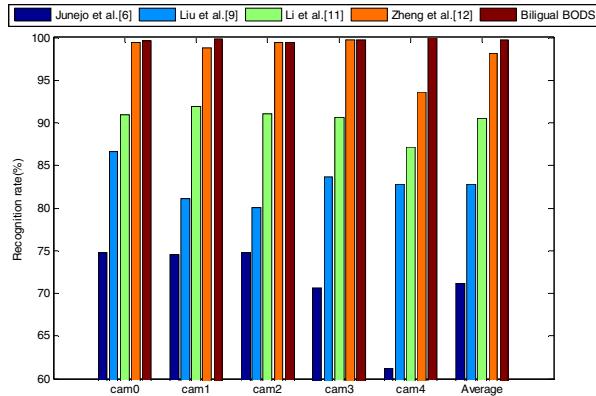


Fig. 2. The result comparision the proposed biligual BODS with other approaches

5 Conclusion

In this paper, we proposed an effective cross-view action recognition algorithm, called biligual BODS. Spatio-temporal patches are extracted as low-level features. The BODS is employed for middle-level representation. Biligual dictionaries are learned by transfer learning, which guarantee that the same action from the two views have same high-level representation. The experimental results on the IXMAS multi-view dataset show the effectiveness of proposed algorithm. What's more important, the proposed biligual BODS can achieve excellent performance on the top view.

Acknowledgements. This work is supported by the NSF of Jiangsu Province BK2010523, the NSF of China under Grant Nos. 61172118 and 61001152, the University Natural Science Research Project of Jiangsu Province under Grant No. 11KJB510012 and the Scientific Research Foundation of Nanjing University of Posts and Telecommunications under Grant No. NY210073.

References

1. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2008)
2. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shap-motion prototype trees. In: Proc. Int'l Conf. on Computer Vision, pp. 444–451 (2009)
3. Gorelick, L., Blank, M., Shechtman, E., et al.: Actions as space-time shapes. IEEE Trans. Pattern Analysis and Machine Intelligence 29(12), 2247–2253 (2007)
4. Grundmann, M., Merier, F., Essa, I.: 3D shape context and distance transform for action recognition. In: Proc. Int'l Conf. on Pattern Recognition, pp. 1–4 (2008)
5. Efros, A., Berg, A.C., Mori, G., et al.: Recognizing action at a distance. In: Proc. Int'l Conf. on Computer Vision (2003)

6. Junejo, I., Dexter, E., Laptev, I., et al.: View-independent action recognition from temporal self-similarities. *IEEE Trans. on Pattern Recognition and Machine Intelligence* 33(1), 173–185 (2011)
7. Farhadi, A., Tabrizi, M., Endres, I., et al.: A latent model of discriminative aspect. In: Proc. Int'l Conf. on Computer Vision, pp. 1–8 (2009)
8. Farhadi, A., Tabrizi, M.K.: Learning to recognize activities from the wrong view point. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I*. LNCS, vol. 5302, pp. 154–166. Springer, Heidelberg (2008)
9. Liu, J., Shah, M., Kuipers, B., et al.: Cross-View Action Recognition via View Knowledge Transfer. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2011)
10. Li, R., Zickler, T.: Discriminative Virtual Views for Cross-View Action Recognition. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2012)
11. Li, B., Camps, O.I., Sznaier, M.: Cross-view activity recognition using Hankelets. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition, pp. 1362–1369 (2012)
12. Zheng, J., Jiang, Z., Phillips, J., et al.: Cross-View Action Recognition via a Transferable Dictionary Pair. In: Proc. of the British Machine Vision Conference (2012)
13. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision* 79(3), 299–318 (2008)
14. Ravichandran, A., Chaudhry, R., Vidal, R.: View-invariant dynamic texture recognition using a bag of dynamical systems. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition, pp. 1651–1657 (2009)
15. Weinland, D., Boyer, E., Ronfard, R.: Action Recognition from Arbitrary Views using 3D Exemplars. In: Proc. Int'l Conf. on Computer Vision (2007)
16. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proc. Int'l Conf. on Computer Vision, pp. 432–439 (2003)
17. Dollar, P., Rabaud, V., Cottrell, G., et al.: Behavior Recognition via Sparse Spatio-Temporal Features. In: 2nd Joint IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
18. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part II*. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
19. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2005)
20. Wang, H., Ullah, M.M., Klaser, A., et al.: Evaluation of local spatio-temporal features for action recognition. In: Proc. British Machine Vision Conference, pp. 127–138 (2009)
21. Doretto, G., Chiuso, A., Soatto, S., Wu, Y.N.: Dynamic textures. *International Journal of Computer Vision* 51(2), 91–109 (2003)
22. De. Cock, K., De. Moor, B.: Subspace angles between linear stochastic models. In: Proc. IEEE Conf. on Decision and Control, pp. 1561–1566 (2000)
23. Jiang, Y., Ngo, C.W.: Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Computer Vision and Image Understanding* 113(3), 405–414 (2008)
24. Aharon, M., Elad, M., Bruckstein, A.: K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* (2006)
25. 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)

Robust Image Set Classification Using Partial Least Squares

Hui Jin¹ and Ruiping Wang²

¹ Peking University, Beijing, 100871, China

² Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
hjin@jdl.ac.cn, wangruiping@ict.ac.cn

Abstract. Image set classification has recently attracted increasing research interest in the field of visual information processing. Different from previous methods that usually characterize set data distribution explicitly using some parametric or non-parametric models, this paper proposes a simple yet effective Partial Least Squares (PLS) regression based method, which seeks to directly learn the underlying statistical relationship between the distributions of set data and their class memberships. With no assumption on the form of set data distribution, the learned model finally reduces to an efficient linear regression from the data space to the class label space, facilitating robust classification of novel test data. Experiments on face recognition and object categorization have shown that the proposed method is competitive to the state-of-the-arts and also quite robust to the noisy set data in practical applications.

Keywords: image set classification, PLS, regression.

1 Introduction

In recent years, with the increase of available video cameras and large capacity storage media, many new applications are emerging, such as visual surveillance, video retrieval, digital photo albums management, etc. In such applications, each object of interest can have a number of image sets for both training and testing, where each image set generally contains lots of images belonging to the same class and covering large appearance variations in pose, lighting, and non-rigid deformations. This is the so-called image set classification problem. By efficiently exploiting the rich set information, more robust object classification can be expected under more realistic conditions [1], [5], [14].

During the past decade, a number of methods have been proposed to solve the problem of image set classification [1], [2], [4], [5]. Generally, these methods make different prior assumptions on the form of set data distribution, and exploit either parametric or non-parametric mathematical models to explicitly characterize data variations in the image set. For parametric modeling, single Gaussian and Gaussian mixture models (GMM) have been explored in earlier works [1], [10] as the parametric distribution function of the image set. Their success highly rely on the assumption that the training and novel test data sets have strong statistical correlations.

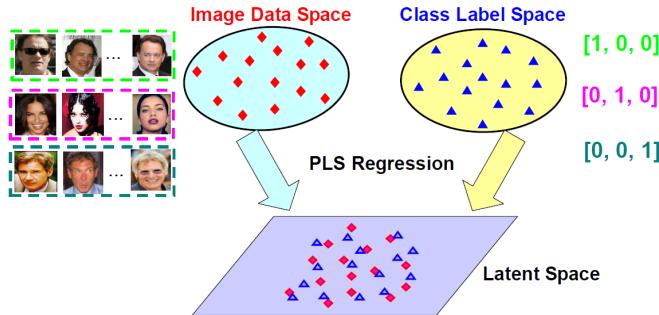


Fig. 1. The basic overview of the proposed SLR method. In the PLS learned latent space, the covariance of the projected image samples and their associated class labels are maximized.

For non-parametric modeling, one class of prevalent methods are based on the model of single linear subspace [5], [15] or more sophisticated nonlinear manifold [6], [12], [14]. Since they can flexibly characterize complex data variation, such methods have gained wide success in past several years. However, linear subspace is a relatively loose representation of the data distribution as noted in [2], while manifold typically needs a large data for reliable estimation, which are unavailable in some practical applications. More recently, a new type of non-parametric methods based on affine subspace model have been introduced [2], [4]. While data variations can be effectively handled, such methods are shown to be sensitive to outliers and have much higher computational cost, due to their inherent single sample-based matching mechanism [13], [14].

In this paper, we propose a simple yet effective Sample-Label Regression (SLR) approach to image set classification. By exploiting the Partial Least Squares (PLS) regression, SLR aims to directly learn the underlying statistical relationship between the set samples distribution and their class labels distribution. Different from previous methods, our approach makes no assumption on the form of set data distribution and the learned model finally reduces to an efficient linear regression from the image data space to the class label space. When applying the learned model to a novel test image set, it only involves computing the class membership score for each image in the set using linear regression, and then aggregating these scores to finally determine the label of the whole set. Fig.1 illustrates the basic idea of the proposed SLR method.

2 PLS-Based Image Set Classification

Formally, given m training image sets as: S_1, S_2, \dots, S_m , we denote $X_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,N_i}]$ ($i=1, 2, \dots, m$) as the data matrix of the i -th image set S_i with N_i samples, where $\mathbf{x}_{i,j} \in \mathbb{R}^d$ is the j -th image sample with d -dimensional feature description. Each set belongs to one of c object classes denoted by $\{L_i \mid L_i \in \{1, 2, \dots, c\}\}_{i=1}^m$. To facilitate following description, we group all training image samples in a single data matrix: $X = [X_1, X_2, \dots, X_m]^T$ of size $N \times d$, where each row of the matrix is an image sample and $N = \sum_{i=1}^m N_i$ is the total number of image samples from all m training image sets.

In the next, we first introduce the basic mathematical model of the Partial Least Squares (PLS) method including both its linear and kernel formulations. Then we elaborate the training and testing framework of exploiting PLS for the specific task of image set classification.

2.1 Background of Partial Least Squares (PLS)

Partial Least Squares (PLS) is a wide class of methods for modeling relations between two sets of observed variables by means of latent variables. In its general form, PLS creates score/latent vectors by using existing correlations between different sets of variables while also keeping most of the variance of both sets. Please refer to [8] for more details.

Let $\mathbf{x} \in \mathcal{X} \subset \Re^d$ denote a d -dimensional vector of predictor variables in the first set of data and similarly $\mathbf{y} \in \mathcal{Y} \subset \Re^c$ denote a c -dimensional vector of response variables from the second set. Observing N data samples from each set of variables, PLS decomposes matrix $\mathbf{X}_{N \times d}$ (it has the same meaning as the above total training data matrix \mathbf{X}) and $\mathbf{Y}_{N \times c}$ into the form

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^T + \mathbf{F}\end{aligned}\tag{1}$$

where \mathbf{T} and \mathbf{U} are $N \times p$ matrices containing the extracted p latent vectors, the $d \times p$ matrix \mathbf{P} and the $c \times p$ matrix \mathbf{Q} represent loadings, and the $N \times d$ matrix \mathbf{E} and the $N \times c$ matrix \mathbf{F} are the residuals. Basically, PLS proceeds to find weight vectors \mathbf{w}, \mathbf{v} such that

$$\max_{\|\mathbf{w}\|=\|\mathbf{v}\|=1} [\text{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{v})]^2 = [\text{cov}(\mathbf{t}, \mathbf{u})]^2,\tag{2}$$

where \mathbf{t} and \mathbf{u} are the column vectors of \mathbf{T} and \mathbf{U} respectively, $\text{cov}(\mathbf{t}, \mathbf{u})$ is the sample covariance. Grouping the sequentially obtained weight vectors \mathbf{w}_i in a matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_p]$, the regression coefficients between the two sets of variables \mathbf{X} and \mathbf{Y} can be estimated by:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{T}^T \mathbf{Y} = \mathbf{X}^T \mathbf{U}(\mathbf{T}^T \mathbf{X} \mathbf{X}^T \mathbf{U})^{-1} \mathbf{T}^T \mathbf{Y},\tag{3}$$

which results in the linear PLS regression $\hat{\mathbf{Y}} = \mathbf{XB}$ [8].

Since it can often bring desirable performance gain by extending linear methods in a so-called RKHS (reproducing kernel Hilbert space) feature space via the kernel trick, in [9] the kernel formulation of PLS (KPLS) has been presented. The basic idea is to map the original \mathcal{X} -space data into a RKHS feature space \mathcal{F} with $\phi: \Re^d \mapsto \mathcal{F}$, where an inner product can be defined using the kernel function as: $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$, and perform the kernel form of the optimization in Eq. (2). Let $\Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)]^T$ be the feature matrix of the training points, the kernel

Gram matrix can thus be written as $\mathbf{K} = \Phi\Phi^T$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Then the regression coefficients \mathbf{B}_ϕ in the feature space will have the form:

$$\mathbf{B}_\phi = \Phi^T \mathbf{U} (\mathbf{T}^T \mathbf{K} \mathbf{U})^{-1} \mathbf{T}^T \mathbf{Y}, \quad (4)$$

Given a testing data example $\mathbf{x}_t \in \Re^d$ in the \mathcal{X} -space, its KPLS prediction \mathbf{y}_t in the \mathcal{Y} -space can be obtained by

$$\mathbf{y}_t^T = [\phi(\mathbf{x}_t)]^T \mathbf{B}_\phi = \mathbf{K}_t^T \mathbf{U} (\mathbf{T}^T \mathbf{K} \mathbf{U})^{-1} \mathbf{T}^T \mathbf{Y}, \quad (5)$$

where $\mathbf{K}_t = [k(\mathbf{x}_1, \mathbf{x}_t), \dots, k(\mathbf{x}_N, \mathbf{x}_t)]^T$.

2.2 Exploiting PLS for Image Set Classification

As illustrated in Fig.1, we exploit PLS regression to directly learn the underlying statistical relationship between the set samples distribution and their class labels distribution. Specifically, we use the training image sets S_i and their associated class labels L_i ($i=1, 2, \dots, m$) to learn the PLS or KPLS latent model. As described in Sec.2.1, the total training data matrix \mathbf{X} of size $N \times d$ acts as the *predictor* matrix $\mathbf{X}_{N \times d}$. For each training image sample $\mathbf{x}_{i,j}$ ($i=1, 2, \dots, m$, $j=1, 2, \dots, N_i$) with its corresponding set class label L_i , we define its class membership indicator vector: $\mathbf{y}_{i,j} = [0, \dots, 1, \dots, 0]^T \in \Re^c$, where the k -th entry being 1 and all other entries being 0 indicates that $\mathbf{x}_{i,j}$ belongs to the k -th class. The *response* matrix $\mathbf{Y}_{N \times c}$ can then be easily constructed with $\mathbf{y}_{i,j}^T$ as its row vector. Taking the two matrices $\mathbf{X}_{N \times d}$ and $\mathbf{Y}_{N \times c}$ as input, either linear PLS or KPLS can then be used to learn the regression model in Eq. (3) or (4) respectively. In the KPLS formulation, we choose the widely used Gaussian RBF kernel function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \sigma^2). \quad (6)$$

In the testing phase, suppose we are given a test image set S_t with $\mathbf{X}_t = [\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \dots, \mathbf{x}_{t,M}]$ as its data matrix containing M image samples, the classification task is to determine the class label of the image set. To this end, for each individual sample $\mathbf{x}_{t,j}$ ($j=1, 2, \dots, M$) we first compute its class membership indicator vector $\mathbf{y}_{t,j}$ (which is c -dimensional) using the PLS/KPLS regression model in Eq. (3) or (4). By aggregating these individual vectors, we then obtain the indicator vector of the whole image set as:

$$\mathbf{y}_t = \frac{1}{M} \sum_{j=1}^M \mathbf{y}_{t,j}. \quad (7)$$

Intuitively, the weight score in the k -th entry of \mathbf{y}_t indicates the probability that the set belongs to the k -th class. Thus, the entry index with the largest score in \mathbf{y}_t finally determines the class label of the test image set S_t .

From above analysis, it can be seen that our method has the following advantages: (1) it makes no assumption of data distribution, thus can be stably applied in different scenarios; (2) it effectively integrates set information from individual samples, resulting in quite robust and efficient classification. Such properties will be verified in the following experiments.

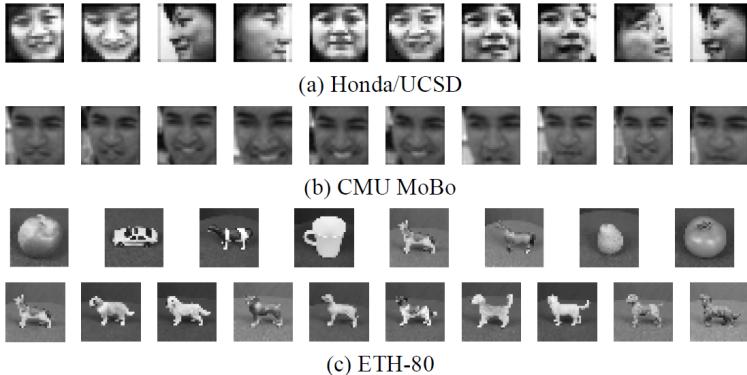


Fig. 2. Example images of the three benchmark databases. In (a) and (b), each row shows representative facial images from one video sequence of an individual. In (c), the first row shows images of the 8 different categories, and the second row shows example images of the 10 different objects for one category.

3 Experiments

We evaluate the proposed method on three widely used datasets: Honda/UCSD [6], CMU MoBo [3] for image sets based face recognition, and ETH-80 [7] for object categorization.

3.1 Databases and Settings

The **Honda/UCSD** consists of 59 video sequences of 20 persons and each video contains about 300~500 frames covering large variations in head pose and facial expression. The **CMU MoBo** contains 96 sequences of 24 subjects and each subject has 4 sequences captured in different walking situations. Each sequence has about 300 frames. We used a cascaded face detector [11] to collect faces in each video, and then resized each face to a 20×20 intensity image. Histogram equalization was used to eliminate lighting effects. Each video generated an image set of faces. The **ETH-80** contains images of 8 categories with each category including 10 objects. Each object has 41 images of different views which form an image set. 20×20 intensity images were also used. Fig. 2 shows some example images from each of the three databases.

For comparison with the literature, we adopted the same protocol as [2],[13]. On all three datasets, we conducted ten-fold cross validation experiments, i.e., 10 randomly selected training/testing combinations, to report average recognition rates of

different methods. Specifically, for both Honda and MoBo, each person had one image set for training and the rest sets for testing. For ETH-80, each category had 5 objects for training and the other 5 objects for testing.

3.2 Comparative Methods and Settings

We compared our approach with several representative non-parametric methods for image set classification, including (i) Mutual Subspace Method (MSM) [15] as the baseline linear subspace based method, and (ii) Affine Hull based Image Set Distance (AHISD) [2], (iii) Convex Hull based Image Set Distance(CHISD) [2], (iv) Sparse Approximated Nearest Point (SANP) [4], which are all affine subspace based methods recently proposed in the literature.

For fair comparison, the key parameters of each method were empirically tuned according to the recommendations in the original references as well as the source codes provided by the original authors. In MSM, PCA was performed to learn the linear subspaces by preserving 95% of data energy. For both AHISD and CHISD, we used their linear version and retained 95% energy by PCA. The error penalty in CHISD was set to $C = 100$ as [2]. For SANP, we adopted the same weight parameters as [4] for the convex optimization.

In our proposed SLR method, we tested both the linear and kernel PLS regression models, referred to as “SLR_L” and “SLR_K” respectively. From Sec.2.1 it can be seen the PLS model has only one parameter, i.e., the number of latent vectors p , which was fixed to 100 in all three databases. We found the classification accuracy was quite stable while varying this number in our experiments. In the KPLS model, there is another important parameter, i.e., the window width σ in the RBF kernel Eq. (6). It was adaptively set for each dataset as the mean of all sample pairs’ Euclidean distances.

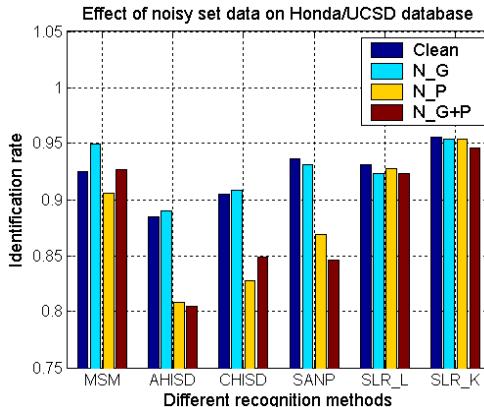
3.3 Results and Analysis

We tabulate the classification results of all methods on the three datasets in Tab. 1. Each reported rate is an average over the ten-fold trials. From the comparison results, we have the following observations: (1) Our method is very competitive to the state-of-the-art ones in all three datasets. In both Honda and ETH-80 datasets, our method delivered the highest rate, and in MoBo dataset, our kernel regression ranked the second highest among all methods. (2) Compared with the linear PLS model, our kernel PLS model can further boost the performance with a modest margin, indicating that the difficult linearly inseparable problem can be effectively alleviated by the nonlinear kernel mapping. (3) In the ETH-80 object dataset, it is interesting to find that the affine subspace methods [2], [4] exhibit much lower accuracy due to that the common intra-class object variations cannot be handled adequately by the single points based matching.

Table 1. Average classification rate of different methods on three datasets by ten-fold trials

Datasets	MSM [15]	AHISD [2]	CHISD [2]	SANP [4]	SLR_L	SLR_K
Honda/UCSD	0.925	0.885	0.905	0.936	0.931	0.956
CMU MoBo	0.852	0.951	0.940	0.963	0.930	0.954
ETH-80	0.878	0.773	0.735	0.755	0.895	0.903

We further conducted experiment to test the robustness of different methods to a practical challenge where the image sets have noisy data from other classes. We followed the same setting as [2] to study this problem on the dataset Honda/UCSD. We tested three cases in which the training gallery and/or the testing probe sets were corrupted by adding one image from each of the other classes. The original clean data and the three noisy cases are referred to as “Clean”, “N_G” (only gallery has noise), “N_P” (only probe has noise), and “N_G+P” (both) respectively. Fig. 3 demonstrates the comparison result. It can be seen that our proposed SLR method shows high robustness against the noisy data challenge, with quite slight accuracy drop. This is mainly attributed to the advantage of our method by effectively integrating the set information from individual samples, without any assumption on the data distribution. Another finding is that the linear subspace based method MSM is more stable than the affine subspace based ones AHISD/CHISD and SANP, since the former treats the set samples as a whole and can suppress the noisy data effectively.

**Fig. 3.** Comparison of different methods on the practical problem of noisy set data

4 Conclusions

In this paper, we have proposed a simple yet effective Sample-Label Regression (SLR) approach to image set classification. With no assumption on the form of set data distribution, our approach exploits the Partial Least Squares (PLS) to directly

learn an efficient linear regression model from the image data space to the class label space. When applying the learned model to classify novel test image set, it involves only simple linear operations and can effectively integrate the set information from individual samples. The extensive experimental results have shown the effectiveness of our method and its favorable robustness to noisy set data in practical applications.

Acknowledgments. This paper is partially supported by Natural Science Foundation of China under contracts No. 61001193 and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract No.4111003.

References

1. Arandjelović, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face Recognition with Image Sets Using Manifold Density Divergence. In: CVPR, pp. 581–588 (2005)
2. Cevikalp, H., Triggs, B.: Face Recognition Based on Image Sets. In: CVPR, pp. 2567–2573 (2010)
3. Gross, R., Shi, J.: The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University (2001)
4. Hu, Y., Mian, A.S., Owens, R.: Sparse Approximated Nearest Points for Image Set Classification. In: CVPR (2011)
5. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. PAMI 29(6), 1005–1018 (2007)
6. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In: CVPR, pp. 313–320 (2003)
7. Leibe, B., Schiele, B.: Analyzing Appearance and Contour Based Methods for Object Categorization. In: CVPR, vol. 2, pp. 409–415 (2003)
8. Rosipal, R., Krämer, N.C.: Overview and Recent Advances in Partial Least Squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) SLSFS 2005. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006)
9. Rosipal, R., Trejo, L.J.: Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. J. Machine Learning Research 2(2), 97–123 (2001)
10. Shakhnarovich, G., Fisher III, J.W., Darrell, T.: Face Recognition from Long-term Observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 851–865. Springer, Heidelberg (2002)
11. Viola, P., Jones, M.: Robust Real-Time Face Detection. Int'l J. Computer Vision 57(2), 137–154 (2004)
12. Wang, R., Chen, X.: Manifold Discriminant Analysis. In: CVPR, pp. 429–436 (2009)
13. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification. In: CVPR, pp. 2496–2503 (2012)
14. Wang, R., Shan, S., Chen, X., Dai, Q., Gao, W.: Manifold-Manifold Distance and Its Application to Face Recognition with Image Sets. IEEE Transactions on Image Processing 21(10), 4466–4479 (2012)
15. Yamaguchi, O., Fukui, K., Maeda, K.: Face Recognition Using Temporal Image Sequence. In: FG, pp. 318–323 (1998)

Research on Quality Improvement of Polarization Imaging in Foggy Conditions

Congli Li, Wenjun Lu, Song Xue, and Yongchang Shi

New Star Research Institute of Applied Technology, Hefei, Anhui China
pbxycip@yahoo.com, wenjun.lu2013@gmail.com, xs_xs6688@sina.com,
756603491@qq.com

Abstract. A new method was presented to improve quality of polarization imaging in foggy weather. In this method, two state-of-art algorithms were used to defog three polarization direction images by polarization imaging. Then Stokes parameter representation was used to parse polarization parameter images. Further image quality assessment based on natural scene statistics was used to verify the acquired polarization parameter images. Sampling images of different foggy density were used in validation experiments. The images were acquired by polarization parameters measurement platform in simulation environment of haze and fog. Subjective and objective assessments show that this method can effectively enhance quality of polarization imaging in foggy conditions. It is easy to be implemented and expanded, and has adaptability to the changes of fog density.

Keywords: polarization imaging, quality improvement, Stokes parameters, defog, image quality assessment.

1 Introduction

In foggy condition, light reflected to the target and surrounding environment is absorbed or scattered as the existence of aerosol particles in the air. This leads to image reduction in contrast, blur and details information loss. With the increasing demand of outdoor vision systems, foggy imaging has become a hot research topic in the field of computer vision. To achieve defogging clarity, polarization imaging is an effective way. It can suppress background clutter, prominent target detail features to enhance the contrast between target and background, which can be separated target from the foggy area [1-2]. Therefore, it has wide application in both civil and military fields.

Foregoing researches [3-4] assume that polarization mainly exists in artificial light and environment light is non-polarization [5]. Schechner [6] considers that natural illuminating light scattered by atmospheric particles is partially polarized and the atmosphere scatter will not change the polarization state of reflex. The methods of polarization filter are based on the characteristics of the polarization of light. Method of polarization filter is applied to imaging long before, which aims to eliminate influence of polarized light. We can eliminate atmospheric scatter to obtain a clear

image by the rotation of the polarizer in front of the camera. However, the influence of fog and haze to imaging cannot be filtered out by only using optical polarization filter in common.

Image defogging method based on polarization filter [7-14] study how to use the atmospheric polarization properties to get scene depth information, then to estimate parameters, and final to restoration.

Y.Wang [12] proposes a polarization dehazing algorithm based on atmosphere background suppression, which makes use of polarization imaging detection system to achieve polarization images from all aspects. P.C. Zhou [13] presents an adaptive image restoration method. By removal of atmospheric light intensity, and compensation for attenuated effect of air light, the intensity information is recovered. Wen.Z.Peng [14] uses edge detection algorithm and the best normal distribution search algorithm to divide the sky, and the atmosphere light intensity and atmosphere transfer coefficient are estimated.

The method emphasizes on intensity image and the degree of polarization (DOP) image. In addition, the quality of the images of these two components will have a direct impact on the defog effect. Therefore, it is necessary to improve quality of polarization imaging and to perceive image quality assessment. Quality assessment of polarization parameters, in particular intensity component and degree of polarization component, is the key point. Research on this direction depends on image acquisition strategy of polarization parameters.

In this paper, quality improvement method of polarization imaging in foggy weather was presented. Increase of defog density is set in simulation environment of haze and fog. The polarization parameters of targets and fog particles are acquired by measurement platform. Innovation in this process is to place defog stage behind calculation of polarization parameters. Subjective observations and objective quality assessment verifies the effectiveness of this method.

Two type of classic defog Tarel [15] algorithm and He [16] algorithm is applied to strengthen defog effect, and NIQE [17] algorithm presented by Mittal [18] in 2013 is used to objective quality assessment of polarization parameters, which is based on natural scene statistics. As its independence to type of image distortion and subjective assessment (MOS, Mean Opinion Score) [19], NIQE algorithm is suitable for assessment of polarization imaging.

2 Quality Improvement Methods of Polarization Imaging

Any target in surface and atmosphere in earth will produce characterized polarization decided by their own properties and basic optical laws in process of reflection, scattering, and transmission. Conventional optical imaging uses light intensity information of object reflection, while polarization imaging uses polarization information of light. Thus, in foggy conditions, polarization imaging has inherent advantages. In the process of polarization imaging, polarization information of the target can be describe as Stokes parameters I , Q , U , V , degree of polarization P , and angle of polarization A . The six parameters are referred as polarization parameter

images. Among polarization parameter images, image I represents the total light intensity, Q represents light intensity difference between components of 0° to 90° , U represents light intensity difference between components of 45° to 135° , V represents light intensity difference between right-rotation and left-rotation circularly polarized component. There are few circularly polarized components in polarization effect on the sun incident on the background and objects of natural atmospheric, V is assumed to zero.

Here we call direct calculation of stokes parameters to acquire degree and angle of polarization as strategy I [12], which is shown in Fig.1 (a). Strategy II presents to defog polarization parameter images respectively after the calculation stage, which is shown in Fig.1 (b). Strategy III presents to defog polarization direction images respectively before calculation of parameters, which is shown in Fig.1 (c).

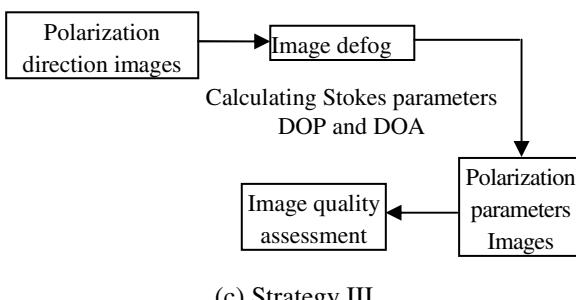
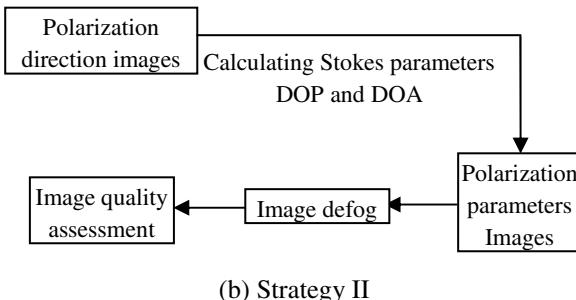
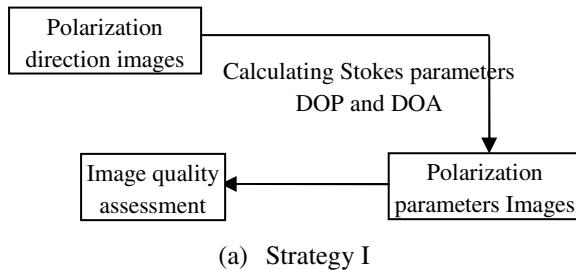


Fig. 1. Three strategy of polarization imaging and assessment of quality improvement

The specific steps of this paper are as follows:

The first step is to select the reference 0^0 and get three polarization direction images of $0^0, 60^0, 120^0$.

In second step Tarel and He algorithms are used to defog the polarization direction images.

The third step is to calculate polarization parameter images before and after defog stage according to following formula respectively.

$$I = \frac{1}{3}(I_{0^\circ} + I_{60^\circ} + I_{120^\circ}) \quad (1)$$

$$Q = \frac{2}{3}(2I_{0^\circ} - I_{60^\circ} - I_{120^\circ}) \quad (2)$$

$$U = \frac{2}{\sqrt{3}}(I_{60^\circ} - I_{120^\circ}) \quad (3)$$

$$P = \frac{\sqrt{Q^2 + U^2}}{I} \quad (4)$$

$$A = \frac{1}{2}ac \tan\left(\frac{U}{Q}\right) \quad (5)$$

The fourth step Tarel and He algorithms are used to defog polarization parameter images respectively.

Finally, subjective observation and NIQE algorithm are used to assess the quality of the five kinds of polarization parameter images. Moreover, through comprehensive comparison and analysis, the method will be verified.

3 Experimental Results and Analysis

In order to verify the feasibility and effectiveness of the proposed method, linear polarization imaging system developed by key laboratory of military opto-electronic is used to acquire the different fog density images to test in simulation environment of haze and fog. The polarization imaging system has the wavelength of 665nm. It uses three pavement CCD imaging modalities, in front of every channel CCD retrofitting a linear polarizer, through the angle between the axis and the selected reference direction, respectively, to $0^0, 60^0$ and 120^0 , obtained polarization image quantized to 8-bit grayscale image. Mist generator used in the laboratory to simulate different concentrations of fog, the use of polarization camera and experimental system for each polarization of the target parameter information.

Experimental environment includes two-color camouflage plate placed in sandy background and place two military vehicles in the ditch. Defog algorithms based on polarization filter are depend on image I and image P .

Due to page limitation, only image I and P in several fog densities is given to contrast, which is shown in Table 1 and Table 2. In two tables, corresponding to each

density, there are five images. First image is acquired by using strategy I, the second image is acquired by using strategy II and Tarel algorithm, the third image is acquired by using strategy III and Tarel algorithm, the forth image is acquired by using strategy II and He algorithm, the last image is acquired by using strategy III and He algorithm.

By subjective observations in Table 1, it can be seen that in all the strategies, quality of image P shows a downward trend with increasing of fog density. In conditions of the same fog density and defog algorithm, strategy III has superior improvement to strategy I and II. In conditions of the same fog density and strategy, He algorithm has superior to Tarel algorithm. This result is also consistent with the conclusions of other researchers [20].

Table 1. Image I comparison of different Strategy and defog algorithms

Strategy Fog density	Strategy I	Strategy II Tarel defog	Strategy III Tarel defog	Strategy II He defog	Strategy III He defog
25%					
30%					
40%					
50%					
60%					
70%					
80%					
90%					

Table 2. Image *P* comparison of different Strategy and defog algorithms

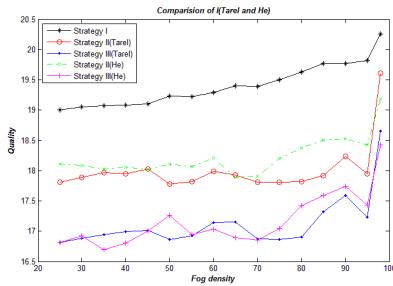
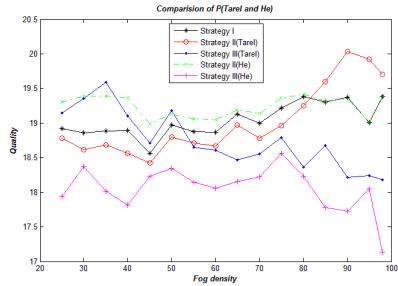
Strategy Fog density	Strategy I	Strategy II Tarel defog	Strategy III Tarel defog	Strategy II He defog	Strategy III He defog
25%					
30%					
40%					
50%					
60%					
70%					
80%					
90%					

Table 3. Percentage of image *I* quality improvement

Fog density	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	98	All	
Strategy	II, Tarel	6.3	6.1	5.7	6.0	5.7	7.5	7.3	6.74	7.6	8.2	8.7	9.2	9.4	7.8	9.5	3.2	7.1
III, Tarel	11.5	11.3	11.2	11.0	11.0	12.3	12.0	11.1	11.6	13.0	13.5	13.9	12.4	11.0	13.1	7.9	11.7	
II, He	4.7	5.1	5.5	5.4	5.7	5.8	6.0	5.6	7.9	7.6	6.7	6.4	6.4	6.3	7.1	5.3	6.1	
III, He	11.5	11.2	12.5	12.0	11.0	10.3	11.8	11.7	13.0	13.1	12.6	11.3	11.0	10.3	12.1	9.0	11.5	

Table 4. Percentage of image P quality improvement

Fog density	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	98	All
Strategy																	
II, Tarel	5.0	4.6	4.0	4.8	3.7	3.9	3.6	3.4	3.1	3.0	2.9	2.8	0.17	1.1	0.3	3.9	3.1
III, Tarel	15.9	14.2	9.5	9.8	3.6	4.0	5.1	4.1	3.8	4.9	4.1	6.8	8.9	9.8	8.2	8.7	7.6
II, He	-1.1	-0.4	0.2	1.0	-1.8	-2.4	-1.1	-2.9	-3.1	-2.5	-0.9	1.8	0.2	-1.4	-0.2	0.3	-0.9
III, He	33.7	39.9	31.9	34.8	34.9	32.7	35.3	32.0	29.2	31.2	29.7	25.4	28.1	27.1	25.5	28.3	31.2

**Fig. 2.** Comparison curve of image I **Fig. 3.** Comparison curve of image P

Further, to verify the effectiveness of proposed method, NIQE algorithm is used to objective image quality assessment. Quality curves of image I and P acquired by different strategy is shown in Fig.2 and Fig.3. Horizontal ordinate indicates fog density (from 25% to 98%). In addition, vertical ordinate indicates quality assessment index, lower the value, higher the quality.

Tables 1 and 2 show that results of objective image assessment is basically followed by subjective observation.

Finally, the percentage of quality improvement of image I and P in using strategy II and III compared to strategy I is shown in Tables 3 and 4.

Tables 3 demonstrates that strategy II and III can improve quality of images I . Tables 4 demonstrates that although strategy II cannot improve quality of image P , strategy III highly improves quality of image P .

Above validations reveal the method proposed in this paper is effective in quality improvement of polarization imaging.

4 Conclusion

In this paper, we propose a new polarization imaging quality improvement method of defog based on polarization filter. The method achieves better quality improvement of key polarization parameters by combining computation of polarization parameter with defog, which contains two presented strategies II and III. Subjective observations and objective assessment experiments show the effectiveness of the method to improve quality of polarization parameters in foggy conditions.

Acknowledgement. This work is supported by the Anhui Natural Science Foundation of China (Grant Nos. 1208085MF97). We thank for Dr. Fan Guo of school of information science and engineering in Central South University to provide complementation file of He algorithms.

References

1. Zhang, X.D., Lin, J.J., Xie, Z., Ji, S., Wu, K.W., Gao, J.: Modeling of Skylight Polarization Pattern Based on Electric Vector. *Acta Electronica Sinica* 38, 2745–2750 (2010)
2. Zhao, Y.Q., Pan, Q., Chen, Y.C., Zhang, H.C.: Clutter Reduction Based on Polarization Imaging Technology and Image Fusion Theory. *Acta Electronica Sinica* 33, 433–435 (2005)
3. Coulson, K.L.: Polarization of light in the natural environment. In: SPIE in Polarization Considerations for Optical Systems, vol. 1166, pp. 2–10. SPIE, San Diego (1989)
4. Quinby-Hunt, M.S., Erskine, L.L., Hunt, A.J.: Polarized light scattering by aerosols in the marine atmospheric boundary layer. *Applied Optics* 36(21), 5168–5184 (1997)
5. Gan, X., Schillders, S.P., Gu, M.: Image enhancement through turbid media under a microscope by use of polarization gating method. *Journal Optical Society of America A* 16(9), 2177–2184 (1999)
6. Schechner, Y.Y., Narasimhan, S.G., Nayar, S.K.: Polarization-based vision through haze. *Applied Optics* 42, 511–525 (2003)
7. Schechner, Y.Y., Karpel, N.: Recovering scenes by polarization analysis. In: MTS/IEEE Oceans, vol. 3, pp. 1255–1261. Marine Technology Society, Kobe (2004)
8. Namer, E., Schechner, Y.Y.: Advanced visibility improvement based on polarization filtered images. In: Polarization Science and Remote Sensing B, vol. 5888, pp. 1–10. SPIE, San Diego (2005)
9. Schechner, Y.Y., Karpel, H.: Recovery of underwater visibility and structure by polarization analysis. *IEEE Journal of Oceanic Engineering* 30(3), 570–587 (2005)
10. Shwartz, S., Namer, E., Schechner, Y.Y.: Blind haze separation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), vol. 2, pp. 1984–1991. IEEE Computer Society, New York (2006)
11. Schechner, Y.Y., Averbach, Y.: Regularized image recovery inscattering media. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(9), 1655–1660 (2007)
12. Wang, Y., Xue, M.G., Huang, Q.C.: Polarization Dehazing Algorithm Based on Atmosphere Background Suppression. *Computer Engineering* 35, 271–275 (2009)
13. Zhou, P.C., Xue, M.G., Zhang, H.K., Han, Y.S., Wang, F.: Automatic image dehaze using polarization filtering. *Journal of Image and Graphics* 16, 1178–1183 (2011)
14. Peng, W.Z.: Polarization dehazing algorithm based on atmosphere scattering model. *Electronic Measurement Technology* 34, 43–45 (2011)
15. Tarel, J.P., Hautiere, N.: Fast Visibility Restoration from a Single Color or Gray Level Image. In: ICCV 2009, pp. 2201–2208 (2009)
16. He, K.M., Sun, J., Tang, X.O.: Single image haze removal using dark channel prior. In: CVPR 2009, pp. 1956–1963 (2009)
17. Mittal, A., Soundarajan, R., Bovik, A.C.: Making a ‘Completely Blind’ image quality analyzer. *IEEE Signal Processing Letters* 20, 209–212 (2013)
18. http://live.ece.utexas.edu/research/quality/niqe_release.zip
19. Jiang, G.Y., Huang, D.J., Wang, X., Yu, M.: Overview on Image Quality Assessment Methods. *Journal of Electronics & Information Technology* 32, 219–226 (2010)
20. Guo, F., Cai, Z.X.: Objective assessment method for the clearness effect of image defogging algorithm. *Acta Automatica Sinica* 38, 1410–1419 (2012)

Human Interaction Recognition by Spatial Structure Models

Jianzhai Wu, Fanglin Chen, and Dewen Hu

College of Mechatronic Engineering and Automation,
National University of Defense Technology, Changsha, Hunan, P.R. China, 410073
dwhu@nudt.edu.cn

Abstract. In this paper, we focus on the recognition and localization of human interactions in real-world videos. It is a difficult challenge because of large variations in person appearance, camera viewpoint, length of video, intra-class variability, and etc. To address these challenges, we present a spatial structure model in this paper. In our model, the crucial movement of each category is represented using a segment of the entire video. To capture the spatial configuration of the human interactions within the video segment, a spatial structure model is built over the segment, and trajectory features are extracted within each cell. The proposed model is trained automatically from real-world videos that are annotated only with the classification label. We examine our approach on the TVHI dataset, which contain 4 complex human interaction action classes. The experimental results demonstrate the effectiveness of our model.

Keywords: Video comprehension, human interaction recognition, spatial structure model, latent SVM.

1 Introduction

Human action recognition in real-world videos is considered an important problem due to the large number of potential applications in areas of visual surveillance, video retrieval and human-computer interfaces. Most of the previous approaches in this field focus on simple human actions, such as “walk”, “run” and “wave hands”, and recognize these actions using a bag-of-word model with various motion descriptors (e.g. [5, 21, 22]). However, in this paper, we focus on the recognition of relatively complex human interactions, such as “handshake” and “hug”. These human action generally have relatively complex spatial configurations; thus, it is difficult to recognize these actions by the widely used bad-of-worlds models.

In this paper, we propose a spatial structure model for recognizing human interactions. Our model is inspired by the spatial models presented in [7] and [10]. The works of [7] and [10] build their models over the entire video sequences. However, in most real-world videos, the total number of frames are not inconsistent, and the occurring time of the key movements that are discriminative for action

categories is unknown in advance. Consequently, we model the occurring time of the key movements as an unknown variable. Then, the action model is built over a video segment rather than the entire video. The proposed model shares some similarity with the temporally deformable part model presented in [12] and the variable-duration hidden Markov model presented in [18]. We adopt the spatial representation because it is more steerable than the models presented in [12] and [18]. First, the proposed representation contains less unknown variables (only one); thus, it is easier to train. Second, our model can be easily extended to include more subcells when fixing the number of unknown variables.

The proposed interaction model is trained based on the latent SVM algorithm presented in [4]. This algorithm proceeds iteratively by alternating between inferring the latent variables and optimizing the model parameters. We employ only one latent variable that represents the start frame of the spatial model. In contrast with [4], we follow the procedure of [12] and use an explicit feature-mapping method [20] to approximate a nonlinear χ^2 kernel. Then, the problem is solved using the LIBLINEAR toolbox [3] that is very efficient for solving linear SVM problems. The proposed model is tested on the TV human interaction (TVHI) dataset [14]. The comparing results validate the effectiveness of the proposed technique.

2 Approach of This Paper

2.1 Spatial Structure Model

In most real-world videos, the total number of frames are not inconsistent, and the occurring time of the key movements that are discriminative for action categories is unknown in advance. Consequently, We build our model on a video segment rather than the entire video. The video segment has a fixed length L and can move temporally within the video sequence, as shown in Fig. 1. The start frame of the temporal segment is unknown in advance. For recognition, only the single segment that provides the largest confidence value is used.

To capture the spatial configuration of the complex actions, the target segment is divided into a set of spatial cells. Fig. 2 presents several different settings of the spatial structure. Then, for each cell, a bag-of-word representation is computed based on dense trajectory-aligned descriptors presented in [22]. Our model is an improved variant of the spatial models presented in [7] and [10]. The works of [7] and [10] build their models over the entire video sequences rather than a temporal video segment.

2.2 Dense Trajectory Aligned Features

To represent the human movements, we make use of the dense trajectory aligned features presented in [22]. The dense trajectory features have been proved to be more efficient than the widely used low-level features obtained by space-time interest point detectors [6] and local descriptors [2, 17].



Fig. 1. Modeling the crucial movements of action categories based on a temporal segment of the video. The start frame of the temporal segment is unknown in advance.

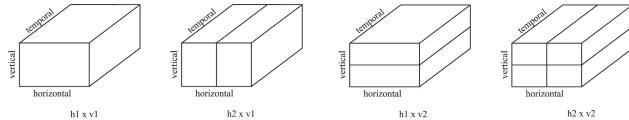


Fig. 2. Example spatial structures tested in this paper

The dense point trajectories are extracted as in [22], and the same parameters are used. The trajectories are represented by four descriptors: the trajectory shape (as a displacement vector), HOG (Histograms of Oriented Gradients [1]), HOF (Histograms of Optical Flow) and MBH (Motion Boundary Histogram [2]). For each of the resultant five descriptor channels (the MBH descriptor has two channels, MBH_x and MBH_y), K-means clustering is used to construct a vocabulary with V words, and the corresponding descriptors are assigned to their closest vocabulary word using the Euclidean distance.

To describe a cell, a bag-of-word histogram (a vector with V dimensions) is computed for each channel. Then, an explicit feature mapping (a vector with $3V$ dimensions) is calculated to approximate a χ^2 kernel [20]. Next, the mapping features of all channels are concatenated together into a unified feature vector (with a total of $15V$ elements). To describe the entire target video segment, we concatenate the feature vectors of all cells within the structural model.

2.3 Training the Model Parameters

To test the effectiveness of the proposed model for encoding the complex human actions, we develop a learning procedure for automatically discovering of the discriminative model representation. The necessary information for training the model, the occurring time of the relevant motions, is unknown in advance.

Our learning procedure is based on the latent SVM framework [4, 23, 25] with a hinge loss function. A linear function is formulated for the model. The latent variable is the temporal location of the fixed-length video segment. The hinge loss object function is minimized using Concave-Convex Procedure (CCCP) [24] which alternates between inferring the latent variables, and optimizing the parameter vector. Once the latent variables are inferred, the parameters are computed by the standard linear SVM, which is implemented using the LIBLINEAR toolbox [3] that is very efficient for solving large-scale linear SVM problems. Considering the large number of possible hypotheses, we adopt a cutting plane method [19] for effectively exploring of the pool. The CCCP process is repeated for several iterations until convergence or a maximum number of iterations is reached.



Fig. 3. Example frames of the human interactions from the TVHI dataset [14]

In most cases, the latent SVM algorithm should be initialized carefully. We choose a simple initialization heuristic. First, we train a standard linear SVM classifier on the bag-of-word representation over the entire video. Then, for every training video, we calculate the confidence of each hypothesis (at every time point of the video) based on the bag-of-word representation over the video segment, and the latent variable is assigned by the hypothesis giving the largest confidence.

3 Experimental Results

3.1 Dataset

We test our method on the TVHI dataset [12] that can be download from the Internet¹. This dataset is composed of 300 video clips compiled from 23 different TV shows. It contains four interactions (Fig. 3): hand shake, high five, hug and kiss, each with 50 videos, as well as 100 negative examples without any of the interactions. The length of the video clips ranges from 30 to 600 frames. The interactions are not temporally aligned. A great degree of variation exists within these clips, such as in the number of actors in each scene, their scales and the camera angle (including abrupt viewpoint changes at shot boundaries).

In our experiments, we follow the split of training/test as in [13]. The performance is evaluated by computing the average precision (AP) for each of the action classes and reporting the mean AP over all classes (mAP) as in [13].

3.2 Testing Different Model Structures

In this subsection, we conduct experiments to demonstrate the influence of different segment lengths and spatial splits of the proposed model. We test four settings of the video segment length: using the entire video, $L=15$, 30 or 60, and examine 4 spatial splits: $h1 \times v1$, $h2 \times v1$ (only splitting horizontally), $h1 \times v2$ (only splitting vertically) and $h2 \times v2$.

The recognition performances of these settings are presented in Table 1. It can be observed that the spatial split of $h2 \times v1$ with a temporal length of 15 frames yields the best result of 53.9%. The settings of length $L=15$ outperforms the use of the entire video, which demonstrates that modeling the crucial movements by latent variables is an effective approach for recognizing these action classes. The reason for the effectiveness of $h2 \times v1$ may be that the TVHI dataset is

¹ http://www.robots.ox.ac.uk/~vgg/data/tv_human_interactions

Table 1. Performances (mAP) of the proposed method with respect to different spatial splits and temporal lengths

Segment-length	h1×v1	h2×v1	h1×v2	h2×v2
entire video	0.449	0.515	0.471	0.493
$L=15$	0.490	0.539	0.478	0.530
$L=30$	0.474	0.504	0.501	0.537
$L=60$	0.449	0.484	0.480	0.527

collected in a relatively controlled setting in which the two actors that perform the interesting interaction movements are generally at the center of the video frames. Consequently, the simple spatial split $h2 \times v1$ succeeds in dividing the two upright actors from each other in most video frames.

3.3 Comparing with the State-of-the-Art Results in the Literature

We compare our approach with several state-of-the-art methods in the literature: the dense trajectory method presented in [22], the spatial pyramid method presented in [7] and [10] and the frame-wise two-person interaction method in [13]. The dense trajectory method [22] obtains the state-of-the-art performance in several data sets containing the KTH [16], YouTube [9], UCF sports [15] and Hollywood2 [11]. The works of [7] and [10] employ the spatial pyramid model from [8] for action recognition in videos. The frame-wise interaction model in [13] combines local and global descriptors in a structured SVM framework, and takes advantage of the visual attention of people by modeling their head orientations.

Table 2 lists the AP results of these methods for each interaction class. For the work of [13], we present only the result based on automatic track that is reported in [13], because the annotation of person bounding box is not used by all the other methods. It can be observed that the proposed method obtains the best average performance of 0.539, which demonstrates the effectiveness of the proposed approach.

Table 2. Comparison with some state-of-the-art methods for TVHI

Method	HS	HF	HG	KS	AVG
Dense trajectory from [22]	0.491	0.546	0.468	0.289	0.449
Frame-wise approach from [13]	0.394	0.458	0.470	0.376	0.424
Spatial pyramid from [7, 10]	0.539	0.566	0.624	0.331	0.515
Ours	0.610	0.571	0.535	0.439	0.539

4 Conclusions

This paper has presented a spatial structure model for recognizing relatively complex human interactions in the real-world video sequences. The proposed model is based a segment of the entire video, and splits the video segment into a set of spatial cells. For each cell, dense trajectory aligned features are extracted to represent the human movements in the local volumes. The model is trained automatically from training videos that are annotated only with the classification label. The proposed model is examined on the TVHI dataset, and the experimental results demonstrate the effectiveness of our method.

Acknowledgments. This work is supported by the National Basic Research Program of China (2013CB329401) and the Natural Science Foundation of China (61203263).

References

- [1] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR (2005)
- [2] Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
- [3] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. (2008)
- [4] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627–1645 (2010)
- [5] Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proc. ICCV (2007)
- [6] Laptev, I., Lindeberg, T.: Space-time interest points. In: Proc. ICCV (2003)
- [7] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. CVPR (2008)
- [8] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
- [9] Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos “in the wild”. In: Proc. CVPR (2009)
- [10] Liu, J., Shah, M.: Learning human actions via information maximization. In: Proc. CVPR (2008)
- [11] Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: Proc. CVPR (2009)
- [12] Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
- [13] Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in tv shows. IEEE Trans. Pattern Anal. Mach. Intell. 34(12), 2441–2453 (2012)
- [14] Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.: High five: Recognising human interactions in tv shows. In: Proc. BMVC (2010)

- [15] Rodriguez, M., Ahmed, J., Shah, M.: Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: Proc. CVPR (2008)
- [16] Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proc. ICPR (2004)
- [17] Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proc. ACM Multimedia (2007)
- [18] Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: Proc. CVPR (2012)
- [19] Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proc. ICML (2004)
- [20] Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: Proc. CVPR (2010)
- [21] Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proc. BMVC (2009)
- [22] Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proc. CVPR (2011)
- [23] Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: Proc. ICML (2009)
- [24] Yuille, A., Rangarajan, A.: The concave-convex procedure (cccp). In: Proc. NIPS, pp. 1033–1040 (2001)
- [25] Zhu, L., Chen, Y., Yuille, A., Freeman, W.: Latent hierarchical structural learning for object detection. In: Proc. CVPR (2010)

Robust Principal Component Analysis for Recognition

Yu Chen and Jian Yang

School of Computer Science and Technology,
Nanjing University of Science and Technology, Nanjing, P.R. China
chenyu1523@gmail.com, csjyang@njust.edu.cn

Abstract. Recently, exactly recovering the intrinsic data structure from highly corrupted observations, which is known as robust principal component analysis (RPCA), has attracted great interest and found many applications in computer vision. Previous work has used RPCA to remove shadows and illuminations from face images. To go further, this paper introduces a method to use RPCA directly for recognition. And the inexact Augmented Lagrange Multiplier algorithm (ALM) is used to solve the RPCA problem. We actually utilize RPCA to reconstruct the testing sample from the training samples and compare the reconstructed one with the original one to do classification. Although the method is not very complicated, through experiments on some face databases we can see that it has better performance compared with some existing methods, especially under rigorous circumstances of occlusions and illuminations.

Keywords: RPCA, recognition, ALM.

1 Introduction

In the world today, massive amounts of high-dimensional data are often obtained in science, engineering, and society. To alleviate the curse of dimensionality and scale, we may leverage on the fact that the data are characterized by low-rank subspaces [1]. That is to say if we stack all the data points as column vectors of a matrix M , the matrix should be low rank: mathematically

$$M = L_0 + S_0 , \quad (1)$$

where L_0 lies in a subspace of low rank and S_0 is the error term. Given a set of training data $M = [m_1, m_2, \dots, m_n]$ with each m_i being generated as in (1), classical Principal Component Analysis (PCA) [2, 3] seeks the best rank-k estimate of L_0 by minimizing the following reconstruction error:

$$\min_U \|M - UU^T M\|_F^2 , \quad \text{s.t. } U^T U = I_k , \quad (2)$$

where I_k is an identity matrix of size $k \times k$, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

PCA gives the optimal estimate when the errors follow Gaussian noise with small variance [4]. However, it breaks down under gross corruption [5]. Even if only one

entry of L_0 is arbitrarily corrupted, the estimated result obtained by classical PCA can be arbitrarily far from the real value. Therefore, Robust PCA should be developed so that a low-rank matrix L_0 can still be efficiently and accurately recovered from a corrupted data matrix.

Wright et al. [5] have proved that: as long as the error matrix S_0 is sufficiently sparse (relative to the rank of L_0), one can exactly recover the low-rank matrix L_0 from $M = L_0 + S_0$ by solving the following convex optimization problem:

$$\begin{aligned} & \text{minimize} && \|L\|_* + \lambda \|S\|_1 \\ & \text{subject to} && L + S = M \end{aligned} \quad (3)$$

The L and S here exactly recover the low-rank L_0 and the sparse S_0 . Lin et al. [6] introduced the Augmented Lagrange Multiplier Method to solve this optimization problem, which is much faster and more accurate compared with other algorithms like the Iterative Thresholding Approach[7,8], the Accelerated Proximal Gradient Approach [9,10,11], and the Dual Approach[12].

In this paper, we try to utilize the RPCA introduced above to do classification. Because we have assumed that the matrix M is low rank despite the noise part, if one face image is quite different from other ones, in the low rank matrix L_0 this face will be changed a lot to become similar to others. Motivated by this, we designed our classifier.

2 ALM for Robust PCA

The method of Augmented Lagrange Multipliers is introduced for solving constrained optimization problems of the kind [13]:

$$\min f(X), \quad \text{subject to } h(X) = 0, \quad (4)$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $h: \mathbb{R}^n \rightarrow \mathbb{R}^m$. We can define the augmented Lagrangian function:

$$G(X, Y, \infty) = f(X) + \langle Y, h(X) \rangle + \frac{\infty}{2} \|h(X)\|_F^2 \quad (5)$$

For the RPCA problem, we can apply the Augmented Lagrange Multiplier method by identifying:

$$X = (L, S), \quad f(X) = \|L\|_* + \lambda \|S\|_1, \quad \text{and } h(X) = M - L - S \quad (6)$$

Then the Lagrangian function is

$$G(L, S, Y, \infty) = \|L\|_* + \lambda \|S\|_1 + \langle Y, M - L - S \rangle + \frac{\infty}{2} \|M - L - S\|_F^2 \quad (7)$$

So the inexact ALM algorithm (updating A and E only once) for solving RPCA problem can be designed by adapting the algorithm of Augmented Lagrange Multipliers [5].

By updating Y in an intelligent manner after each minimization of the form, the algorithm converges to the exact optimal solution, even without requiring ∞ approaching infinity [13].

3 Classifier Design

Previous work has used RPCA to separate targets from background and to remove shadows and illuminations from face images [4]. To go further, we consider if one column of M does not even come from the same class with the others. Fig.1 shows an experiment to explain our idea. We choose one image from AR face database as a testing sample and conduct three groups of experiments. Firstly, we stack the testing sample with three images from the same subject as column vectors of the matrix M in the RPCA problem (3). In the next experiments, we stack the testing sample with three images from different subjects. In the low rank part, we can see the fourth face image is recovered to be similar to the first three images in M . If it comes from the same class with the resting ones, it won't change a lot. If it comes from a different class with the others, it will change a lot to become similar to others in the low part.

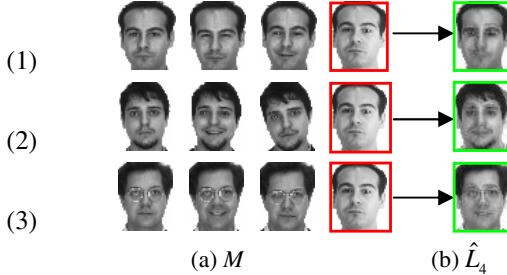


Fig. 1. Three groups of experiments to show our idea. (a) Four images stacked as column vectors of M . (b) The fourth column of the low rank approximation L recovered by inexact ALM.

This is because in the model we have assumed that the matrix L recovered from M is low rank. That is to say the column vectors of L are linear dependent. So the image which comes from the different class with others has to be reconstructed to be similar to others to satisfy this assumption. Based on this, by computing the distance between the original sample and the reconstructed one, we can design our classifier.

Given training samples of n classes, each class has k samples:

$$\begin{aligned} & X_{11}, X_{12}, X_{13} \dots X_{1K}; \\ & X_{21}, X_{22}, X_{23} \dots X_{2K}; \\ & X_{n1}, X_{n2}, X_{n3} \dots X_{nK}; \end{aligned} \quad (8)$$

where X_{ij} is a column vector. For any testing sample y , we construct:

$$M_i = [X_{i1}, X_{i2}, X_{i3} \dots X_{iK}, y]. \quad (9)$$

Then we use the inexact ALM to recover the low rank approximation L_i and the sparse part S_i of M_i (we use the default value of parameter λ in RPCA):

$$\begin{aligned} L_i &= [L_{i1}, L_{i2}, L_{i3} \cdots L_{iK}, y_{iL}], \\ S_i &= [S_{i1}, S_{i2}, S_{i3} \cdots S_{iK}, S_{iL}], \end{aligned} \quad (10)$$

where $M_i = L_i + S_i$. We may define y_{iL} as the image of y in training samples of class i . For all $i \in [1, 2, 3 \cdots n]$, we do this decomposition. So we get the images of y in training samples of all classes:

$$y_{1L}, y_{2L}, y_{3L} \cdots y_{nL}. \quad (11)$$

At last, we make decisions by:

$$\begin{aligned} \text{if } \min_{i \in [1, 2, 3 \cdots n]} \|y - y_{iA}\|_* &= \|y - y_{jA}\|_*, \\ \text{then } y &\in \text{class } j. \end{aligned} \quad (12)$$

Let $\|M\|_* \doteq \sum_i \sigma_i(M)$ denote the nuclear norm of the matrix M , i.e. the sum of the singular values of M .

4 Experiments

We test our method for recognition on face databases, including the Yale database, the extended Yale B database [14] and the AR database [15].

4.1 Experiments on the Yale Database

The Yale database consists of 15 subjects. Each subject has 11 images under different illuminations and expressions. We cropped each image to the size of 100×80 . Training samples change from the first three to the first six images and we use the rest for testing. We compare our method with Eigenface plus NN classifier, Fisherface plus NN classifier, CRC [16], LRC [17] and SRC [18].

We can see from Fig.2 that our method has good performance but is a little weaker than CRC and LRC. We will talk about this phenomenon in section 4.3.

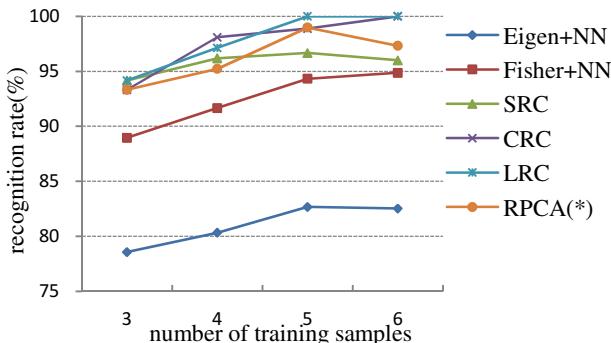


Fig. 2. Experiments on Yale database. RPCA classifier is compared with some classifiers.

4.2 Experiments on the Extended Yale B Database

The extended Yale B face database [14] consists of 2432 images from 38 subjects. Each subject has 64 images with illuminations from 64 different directions. We cropped each image to the size of 96×84 . There are five subsets in extended Yale B database. The condition of illuminations becomes more rigorous from Subset 1 to Subset 5. We choose images in Subset 1 for training and the left 4 subsets for testing respectively. Our method is also compared with other classifiers introduced above. We can see that RPCA always shows better performance especially when the condition of illuminations is rigorous.

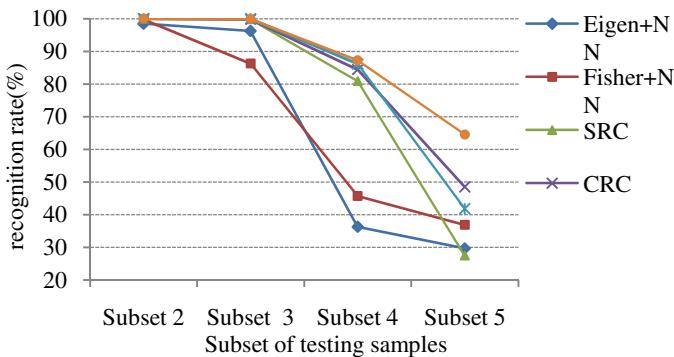


Fig. 3. The recognition rates (%) of Eigen+NN, Fisher+NN, SRC, LRC, CRC and the proposed RPCA classifier on the extended Yale B database

4.3 Experiments on the AR Database

The AR face database [15] consists of images from 126(70 male, 56 female) subjects. Images of 120 subjects from this database were separated into two sessions. Session 2 was collected two weeks later after session 1. We manually cropped the face portion of the image and then normalized it to 50×40 pixels.

Fig.4 shows a group of samples from one subject. At first, we consider the effect of illuminations. We choose images (1), (2), (3), (4) for training and (5), (6), (7) for testing. Our method is a little weaker than SRC and CRC. From experiments in section 4.2 we can tell our method shows obvious superiority only when the influence of illuminations in the testing samples is very obvious. So the condition of illuminations here in the testing sample is not rigorous enough to show the advantage of our method under large noise. On the other hand, the training samples in this experiment are affected by expressions. The expression of a face cannot be expressed by a sparse matrix directly while the illumination and occlusion can. This is why our method is weaker in this experiment.

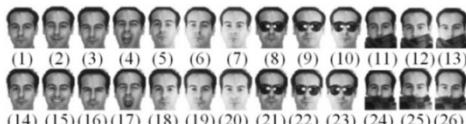


Fig. 4. All 26 face images of one subject from the AR database

Table 1. The recognition rates (%) of Eigenface+NN, Fisherface+NN, SRC, LRC, CRC and the proposed RPCA classifier of experiments for illuminations on the AR database.

Eigen+NN	Fisher+NN	SRC	LRC	CRC	RPCA
87.5	89.4	98.6	89.7	98.3	95.3

Table 2. The recognition rates (%) of Eigenface+NN, Fisherface+NN, SRC, LRC, CRC and the proposed RPCA classifier of experiments for occlusions on the AR database.

Eigen+NN	Fisher+NN	SRC	LRC	CRC	RPCA
67.8	68	74.5	69.2	74.1	80.4

Next, we consider the effect of occlusions. Firstly, we choose images (1), (2), (3), (5), (6), (7) for training and (8)-(13) for testing. Then, we do the same experiment on session 2. We average the results of these two experiments and show it in Table.2.

From Table.2, we can find our classifier has the best performance. In the previous experiments, we have also seen that our method shows advantages when the noise in the image is large. This is because RPCA has the ability to remove large noise into the sparse part S_0 , as we show in Fig.5. Supposing we have some training samples without occlusions and a testing sample with occlusions, the occlusion on the testing sample will be considered as noise and be removed into the sparse part to satisfy the condition that L_0 is low rank. As long as the noise on faces can be removed into the sparse matrix in the form of $S_0 = M - L_0$, our classifier will have good performance. This is why we have good results under situations of illuminations and occlusions.

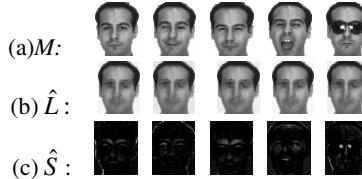


Fig. 5. The ability of RPCA classifier to remove occlusions. (a) Original images from AR database. (b) Low-rank part recovered by inexact ALM. (c) Sparse part recovered by inexact ALM.

5 Conclusions

We have introduced a method to utilize RPCA for recognition. Our idea is to reconstruct the testing sample from the training samples and compute the distance between the reconstructed one and the original one to do classification. Through experiments we can see our method is comparative to other methods under normal conditions and our method is better under rigorous circumstances of illuminations and occlusions.

References

1. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. *Psychometrika* 1, 211–218 (1936)
2. Hotelling, H.: Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417–441 (1933)
3. Jolliffe, I.: *Principal Component Analysis*. Springer (1986)
4. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(9), 1546–1562 (2007)
5. Candes, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *J. ACM* 58, 11:1–11:37 (2011)
6. Lin, Z., Chen, M., Ma, Y.: The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrix. Technical Report UILU-ENG-09-2215, UIUC, arXiv: 1009.5055 (2009)
7. Yin, W., Hale, E., Zhang, Y.: Fixed-point continuation for ℓ^1 -minimization: Methodology and convergence (2008) (preprint)
8. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ^1 -minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences* 1(1), 143–168 (2008)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1), 183–202 (2009)
10. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. *SIAM Journal on Optimization* (2008) (submitted)
11. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* 27(2), 372–376 (1983)
12. Ganesh, A., Lin, Z., Wright, J., Wu, L., Chen, M., Ma, Y.: Fast Algorithms for Recovering a Corrupted Low-Rank Matrix. In: International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (2009)
13. Bertsekas, D.: *Constrained Optimization and Lagrange Multiplier Method*. Academic Press (1982)
14. Lee, K.C., Ho, J., Driegman, D.: Acquiring Linear Subspaces for Face Recognition under Variable Lighting. *IEEE Trans. PAMI* 27(5), 684–698 (2005)
15. Martinez, A.M., Benavente, R.: The AR Face Database. CVC Technical Report #24 (1998)
16. Zhang, L., Yang, M., Feng, X.C.: Sparse representation or collaborative representation which helps face recognition? In: ICCV (2011)
17. Naseem, I., Togneri, R., Bennamoun, M.: Linear Regression for Face Recognition. *IEEE Trans. PAMI* 32(11), 2106–2112 (2010)
18. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. PAMI* 31(2), 210–227 (2009)

Time-Varying Distributed Resource Allocation Based on Thermal Minority Game

Jin Liu, Qianping Wang, Zhizhen Liang, and Wei Chen

School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, Jiangshu 221116, China
{liujincumt,davior.chen}@gmail.com,
{qpwang,liang}@cumt.edu.cn

Abstract. In this paper, a thermal minority game (TMG)-based distributed resource allocation scheme is presented. By taking advantage of the continuous bid property of TMG and after some improvements, the modified TMG-based scheme we proposed herein can work well when the resources are divisible and capacity is dynamic. For aims of comparison, a number of experiments are carried out, and the experimental results demonstrated the superior performance of the proposed scheme.

Keywords: game theory, time-varying channels, distributed resource allocation, fair allocation.

1 Introduction

Distributed resource allocation have been widely studied across different disciplines ranging from economy, computer science, etc. [1][2][3]. The resources to be shared can be energy, information, materials, bandwidth and so on. The objective of the research issue is to figure out a solution which can satisfy most of the demands and make good use of the resources. Taking communication networks as an example, bandwidths are allocated to each user to communicate with each other. Given the total resources in the system, how to allocate the bandwidth to each Source Destination(SD) pair in an efficient and fair way can be considered as a kind of distributed resource allocation problem.

Game theory based methods have been proposed to model distributed resource allocation problem between users [4][5][6], and minority game based methods[7][8]. In [7], an extension of minority game is presented in which each player is connected with its neighbors, and the strategy of each user is influenced by decisions from its neighbors. In [9], a discussion on distributed resource allocation with time-varying resource capacity is presented. The authors model the research issue as a minority game.

In this paper, a novel scheme of Thermal Minority Game (TMG)-based distributed resource allocation is presented. By taking advantage of the property of continuous bid in TMG, the action of each user is a continuous value rather than

a binary value, which is more suitable for cases where the resources are divisible. The proposed scheme is validated through experimentations, and the results have demonstrated that the system could converge to the resource capacity at a quick rate and result in a comparably fair allocation.

2 Distributed Resource Allocation Based on Modified TMG

Taking the continuous bid property of TMG into consideration, it is intuitive to model resource allocation for divisible goods as a TMG. But the mapping is not direct and the original model of TMG requires modifications. Firstly, as is shown in [10], the cooperation state of TMG is with the property of $E[A(t)] = 0$. While in real world problems, the expectation of total attendance $E[(t)]$ is usually a positive real number C . Secondly, users in the system should be able to perceive the change of resource capacity and consequently follow the change. As an effort for amending the challenges listed above, we introduced several modifications into TMG. Firstly, the negative part of strategy space is used instead of the whole strategy. Secondly, instead of the random information, a unit vector is used as the common information shared for users for action generation. Last but not least, the update of the virtual value in each strategy set is calculated based on a function involves the product between individual's action with the total attendance, rather than use the product directly. When the user is in the minority side, the virtual value of the strategy would be increased by one, and decreased otherwise.

As is aforementioned, it is only the negative half of the d -dimension real strategy space \mathbf{R}^d used rather than the whole strategy space. By this means, the whole strategy space is a hyper-sphere on which each strategy is a d -dimensional point with $|\vec{r}| = 1$, and the negative part of $-\mathbf{R}^d$ takes $1/(2^d)$ of the whole strategy hyper-sphere. Users pick s strategies randomly from $-\mathbf{R}^d$, in this sense, users in the system are heterogenous.

The history information of each agent $\vec{\mu}(t)$ in this game in round t is in the form

$$\vec{\mu}(t) = \{\mu(t-1), \mu(t-2), \dots, \mu(t-d)\} \quad (1)$$

where the elements in the memory vector $\mu(t-i), i = \{1, \dots, d\}$ is composed of the history states of the resource in round $t-i$.

And the history information $\vec{I}(t)$ in round t observed by each agent is formulated as

$$\vec{I}(t) = \left\{ \frac{\mu(t-1)}{\mu}, \frac{\mu(t-2)}{\mu}, \dots, \frac{\mu(t-d)}{\mu} \right\} \quad (2)$$

where

$$\mu = \left(\sum_{i=1}^d \mu^2(t-i) \right)^{1/2} \quad (3)$$

After assigned strategies and formed the unit vector \vec{r}_i^* , users would pick one strategy \vec{r}_i^* out of its s strategies, and calculate its bid $b_i(t)$ for next round

$$b_i(t+1) = \vec{r}_i^*(t) \cdot \vec{I}(t) \quad (4)$$

After the agent i gets its bid for the next round, the action $a_i(t+1)$ of the agent for the next round can be computed by summing all of the bids in the previous iterations

$$a_i(t+1) = \sum_{j=0}^t b_i(t-j) \quad (5)$$

After each agent took action and submitted its action to the resources. The resources will then sum all the submitted actions together, and minus the upper bound of the resource it can provide and broadcast the result $\mu^j(t)$ back to each user

$$\mu^m(t) = \sum_{i=1}^N a_i^m(t) - C_m \quad (6)$$

where C_m is the capacity of the resource m , $\mu^j(t)$ stands for the winning choice or history information in MG.

At the end of each round, users would calculate the virtual value $p_i^s(t+1)$ for their strategies for the next round. The update for p_i^s is in a similar way with that in the original TMG,

$$p_i^s(t+1) = p_i^s(t) - update_i^s(t) \quad (7)$$

and the update is defined as

$$update_i^s(t) = \begin{cases} -1 & \text{if } \mu(t) * b_i(t) > 0 \\ 0 & \text{if } \mu(t) * b_i(t) = 0 \\ 1 & \text{if } \mu(t) * b_i(t) < 0 \end{cases}$$

The update of virtual value listed above shares common indicates with virtual value update in other MG methods. For instances, when the resource in last round is underused ($\mu(t) < 0$) and meanwhile the user required more resources in last round ($b_i(t) > 0$), the product between them would result in negative. Hence the $p_i^s(t+1)$ would be increased by 1 compared $p_i^s(t)$, which indicates that the strategy s predicted the action correctly, and it would be granted with higher chances for the following round.

The reason that the proposed TMG-based distributed resource allocation is adaptive to the capacity change can be explained as the follows. When the resource is overused for more than d rounds, the elements in $\vec{I}(t)$ would become positive, and hence the inner product between $\vec{I}(t)$ and \vec{r}_i^* would become negative. Namely, the bid for the next round $b_i(t+1)$ would become negative, which indicates that the user i would release its usage on this resource by $|b_i(t)|$. Vice versa, when the resource is underused, users would increase their usages. As illustrated previously, the $b_i(t+1)$ stands for the increment or decrement of resource

usage between iterations. And by taking the cumulated bids as the action for each user, users could identify the trend based on the price information, and thus adjust their usages accordingly until the total usage reaches the capacity. That is the reason why the system bounds around the optimal usage of the resources.

For the next iteration, the process will repeat with the latest virtue values and memories, and the strategy with the highest probability will be selected, and bids and actions are calculated by users and the process goes on.

The procedure of the proposed TMG-based distributed resource allocation can be summarized as the follows.

1. Initialize user i , $i = 1, \dots, N$ with strategies r_i^s , $s = 1, \dots, s$ and memory information $I(t)$;
2. Pick strategy $r_i^*(t)$ with the highest virtual value for each user u_i ;
3. Users generate bid $b(t+1)$ based on the memory $I(t)$ and the selected strategy $r_i^*(t)$ according to Eq. (4);
4. Users calculate their resources by summing together their history bids according to Eq. (5);
5. Resources updates the states and broadcast back the price signal to users according to Eq. (6);
6. Users update the memory $I(t + 1)$ based on the observation of price signal according to Eq. (2);
7. Users update the virtue value $p_i^s(t + 1)$ for the strategy according to Eq. (7);
8. Check the convergence of the algorithm. The algorithm stops if the convergence is reached and goes to step 2 the otherwise.

3 Simulation Study

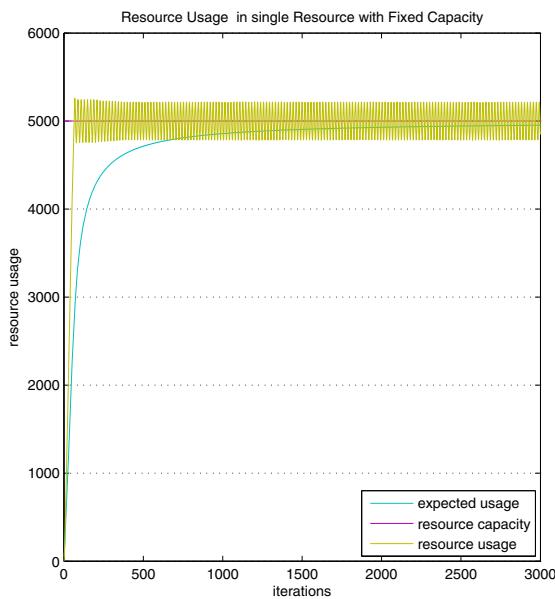
To evaluate the proposed TMG-based distributed resource allocation, we conducted a number of simulations with both single resource and multiple resources, with fixed and time-varying capacities. The parameters settings are $d=10$, $s=2$, $N=100$, $iteration=3000$. We investigate the performance from both the system and user1 point of view with fixed and time-varying resource capacity.

3.1 Fixed Resource Channel

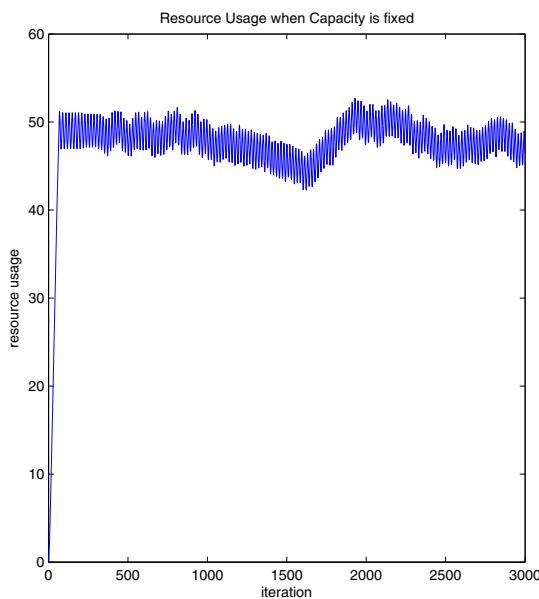
In the fixed resource channel case, the capacity of resource is set to be 5000, and the usage of the system is shown as in Figure 1(a), from which it can be observed that the total usage of the resource reaches the capacity within a few of iterations.

Figure 1(a) shows the resource usage of user1 in fixed capacity case, from which it can be observed that the resource occupied by user1 is around 50, which is the average amount of resources shared by all the users $C/N = 5000/100 = 50$. It can also be seen that the usage of user1 does not strictly equal to 50, rather than that, the usage fluctuates around the average.

Table 1 shows that resource usage among 5 users in the system with single resource of fixed capacity, from which it can be seen that users are sharing the resource almost equally, which is quite a fair means.

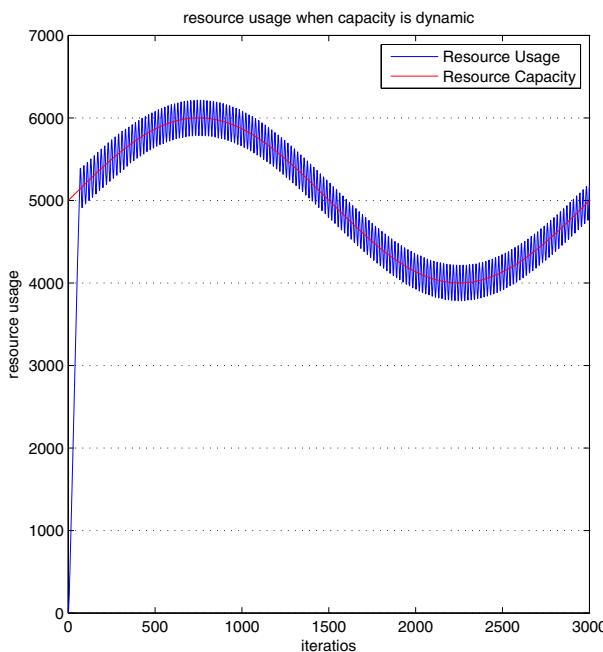


(a) Total usage

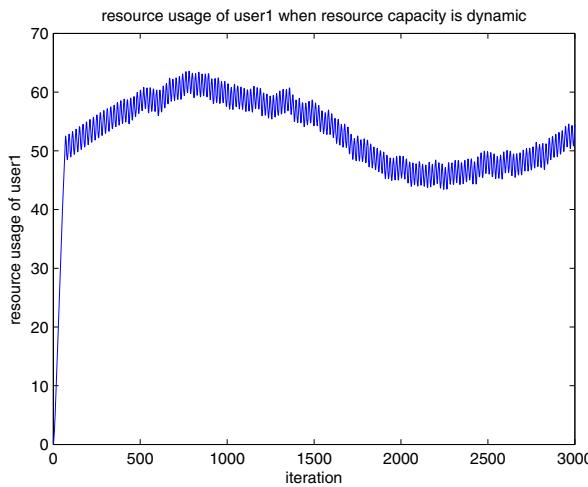


(b) Single user

Fig. 1. Usage of single resource with fixed resource capacity



(a) Total usage



(b) Single user

Fig. 2. Resource usage when capacity is time-varying

Table 1. Resource usage among users in single fixed resource capacity 5000

	10th iteration	20th iteration	100th iteration	1000th iteration	2000th iteration	expected usage
User1	5.53216	13.6918	47.222	48.5182	49.832	47.4752
User2	6.67646	13.7945	40.1526	39.7899	42.6414	40.5861
User3	4.52189	12.7202	45.9594	47.8539	45.7547	46.2876
User4	6.06112	13.3108	40.2885	40.7753	43.2599	41.2484
User5	7.21482	16.3984	51.8317	54.2167	54.4547	53.6403

3.2 Resource Allocation with Time-Varying Capacity

Figure 2(a) shows the system resource usage when capacity is time-varying, in which the resource is changing against function $capacity = 5000 + 1000 * \sin(t * 2\pi/T)$, where T is the iteration counts. Based on the Figure, it can be observed that the proposed scheme is capable to adapt to the capacity in an efficient way.

Figure 2(b) shows the resource usage of a single user in dynamic resource capacity case with the same parameter settings. From the figure, it can be clearly seen that although the resource usage of user1 does not strictly equal to the capacity change, the general trend of the resource usage of the user is consistent with the capacity change.

4 Conclusion

In this paper, a TMG-based distributed resource allocation scheme is presented. Being different from previous work, resource usage of each user in the proposed scheme is a continuous value rather than a binary bid. Hence the state of the resource is determined by the total usage over all the users rather than by the number of users, which is closer to the real world scenarios such as communication networks or share market trading, etc. Meanwhile, by introducing a couple of modifications, users in the presented scheme can identify the trend of the capacity changes, and increase or decrease its resource occupancy accordingly to follow the trend. The proposed scheme is evaluated under both single and multiple resources scenario, with fixed and dynamic resource capacities. The experimental results demonstrated that the proposed scheme is capable to allocate the resources in an efficient and fair way, and hence is suitable to model real world problems with strategic users, such as in bandwidth allocation and spectrum allocation in communication networks.

Acknowledgments. The paper is supported by the National Natural Science Foundation of China (Grant No. 61303182), the Natural Science Foundation of Jiangsu Province (Grant No. BK20130210), Fundamental Research Funds for the Central Universities (Grant No. 2012QN17), the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20120095120026), the Postdoctoral Science Foundation of China (Grant No. 2012M521144), the Postdoctoral Science Foundation of Jiangsu Province (Grant No. 1301120C).

References

1. Johari, R., Tsitsiklis, J.N.: A scalable network resource allocation mechanism with bounded efficiency loss. *IEEE Journal on Selected Areas in Communications* 24(5), 992–999 (2006)
2. Ganesh, A., Laevens, K., Steinberg, R.: Congestion Pricing and Noncooperative Games in Communication Networks. *Operation Research* 55(3), 430–438 (2007)
3. Jain, R., Walrand, J.: An efficient nash-implementation mechanism for network resource allocation. *Automatica* 46(8), 1276–1283 (2010)
4. Cong, L., Zhao, L., Zhang, H., Yang, K., Zhang, G., Zhu, W.: Pricing-based game for spectrum allocation in multi-relay cooperative transmission networks. *IET Communications* 5(4), 563–573 (2011)
5. Zhang, Z., Shi, J., Chen, H., Guizani, M., Qiu, P.: A cooperation strategy based on nash bargaining solution in cooperative relay networks. *IEEE Transactions on Vehicular Technology* 57(4), 2570–2577 (2008)
6. Zhu, H., Liu, K.J.R.: Noncooperative power-control game and throughput game over wireless networks. *IEEE Transactions on Communications* 53(10), 1625–1629 (2005)
7. Galstyan, A., Czajkowski, K., Lerman, K.: Resource allocation in the grid with learning agents. *Journal of Grid Computing* 3, 91–100 (2005)
8. Chow, F.K., Chau, H.F.: Multichoice minority game: dynamics and global cooperation. *Physica A: Statistical and Theoretical Physics* 337(1-2), 288–306 (2004)
9. She, Y., Leung, H.-F.: An adaptive strategy for resource allocation with changing capacities. In: Zhou, J. (ed.) *Complex 2009. LNICST*, vol. 5, pp. 1410–1423. Springer, Heidelberg (2009)
10. Cavagna, A., Garrahan, J.P., Giardina, I., Sherrington, D.: Thermal model for adaptive competition in a market. *Phys. Rev. Lett.* 83(21), 4429–4432 (1999)

Salient Object Detection via Fast Iterative Truncated Nuclear Norm Recovery

Chuhang Zou, Yao Hu, Deng Cai, and Xiaofei He

Zhejiang University, HangZhou, China

{zouchuhang, huyao001, dengcai, xiaofeihe}@gmail.com

Abstract. Salient object detection is a challenging problem in many areas such as image segmentation and object recognition. Many approaches reveal that the background of an image usually lies in a low-dimensional subspace, while the salient regions perform as noises. Conventional methods apply nuclear norm minimization to recover the low-rank background to get the saliency. However, the nuclear norm could not approximate the rank operator properly. In this paper, we propose a novel salient object detection method called Fast Iterative Truncated Nuclear Norm Recovery (FIT) to detect salient objects. Recent proposed Truncated Nuclear Norm is used as a convex relaxation of the rank operator, which consequently guarantees a higher accuracy while reducing time consumption in saliency detection. Series of experiments have been conducted on widely used public database. The results demonstrate the efficiency of our proposed algorithm compared with the state-of-the-art.

Keywords: Salient Object Detection, Low-rank Matrix Recovery, Truncated Nuclear Norm.

1 Introduction

Generally, the goal of salient object detection is to determine the region of an object that is salient visually, which is a fundamental research problem arising in areas such as content-based retrieval [3] and image segmentation [8].

Many researchers propose to detect saliency based on spatial contrast that focus on local or global contrasts among image features [7,4]. Others apply frequency domains analysis methods to extract multiscale features based on the spectrums of images [5,9]. The above methods require the target image to have a high quality and cannot be applied in the noisy case.

Recent work [12] has revealed that the salient object detection can be categorized into low-rank matrix recovery problem in a certain feature space. This is because the background of an image in the feature space usually lies in a low-dimensional subspace, while the rest salient regions can be seen as noises or errors (probably sparse), which is illustrated in Fig.1. Traditional techniques use nuclear norm based heuristics such as Robust Principle Component Analysis (RPCA) to recover a low-rank matrix with sparse noise. However, consider the rank function in which all the non-zero singular values have equal contributions,

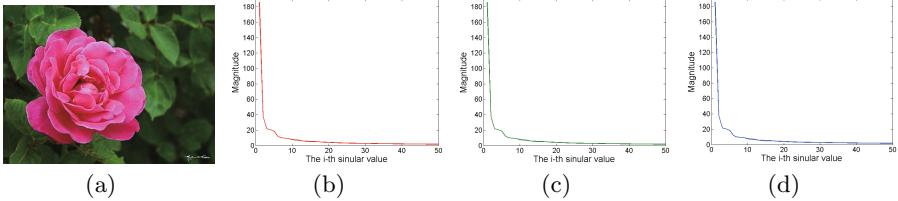


Fig. 1. (a) A 400×300 image. (b) The red channel singular value of the background of the image. (c) The green channel singular value of the background of the image. (d) The blue channel singular value of the background of the image. The ground truth background is dominated by several first singular values, and is obviously low-rank.

the nuclear norm treats the singular values differently by adding them together. Therefore, nuclear norm cannot approximate the rank operator properly and results in an unsatisfactory detection performance.

In this paper, we propose a novel salient object detection method called Fast Iterative Truncated Nuclear Norm Recovery (FIT) to detect salient object accurately and efficiently. Different from previous approaches that apply RPCA for low-rank matrix recovery, we utilize *Truncated Nuclear Norm* (TNN) [6] to capture the low-rank structure of the background more accurately. Since TNN is non-convex, an efficient iterative scheme is applied to solve the final optimization problem. Experimental results on public dataset show a better performance of our methods both on accuracy and efficiency compared with the state-of-the-art.

The remainder of the paper is organized as follows. In section 2, we briefly review the related work about saliency detection. We then present our proposed FIT method in section 3. The extensive experiment results are shown in Section 4. Finally, we provide some concluding remarks in Section 5.

Notions: Given a matrix $X \in \mathbb{R}^{m \times n}$, σ_i is the i -th largest singular value of X , the nuclear norm is defined as: $\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i$. Let $X = U\Sigma V^T$ be the singular value decomposition for X , where $\Sigma = \text{diag}(\sigma_i)$, $1 \leq i \leq \min\{m, n\}$, the “shrinkage” operator $D_\tau(X)$ is defined as $D_\tau(X) = U\Sigma_\tau V^T$.

2 Related Work

Many efforts have been devoted to salient object detection. One solution is to apply contrast-based features such as colors [13], local maxima of activities [7], global contrast and spatial coherence [4] to measure the differences of a region from others. Another kind of solution is frequency domain analysis. Applying amplitude spectrum [5] and phase spectrum [9] are two typical methods. Above methods seek a deep representation of visual saliency to imitate the real human visual system, but are not perfect enough to be robust to noises.

To handle with the noisy cases, several approaches categorize saliency detection problem into recovering a low-rank matrix with sparse noises. Given the feature matrix F , F is assumed to consist of two parts: $F = X + E$, where X is

the low-rank matrix and E is a sparse matrix representing as saliency. The core issue for extracting the saliency from the feature is to recover both the low-rank and the sparse matrix, shown as follows:

$$(X^*, E^*) = \arg \min_{\substack{X, E \\ s.t.}} \text{rank}(X) + \lambda \|E\|_0 \quad (1)$$

$$F = X + E.$$

Problem (1) is NP-hard because of the non-convex and non-smooth nature of the rank operator $\text{rank}(\cdot)$ and $\|\cdot\|_0$. Based on the recent development for low rank matrix recovery, we alternatively solve the convex surrogate as follows:

$$(X^*, E^*) = \arg \min_{\substack{X, E \\ s.t.}} \|X\|_* + \lambda \|E\|_1 \quad (2)$$

$$F = X + E.$$

Recent approaches propose to solve problem (2) under the scheme of Robust Principle Component (RPCA). However, since nuclear norm cannot approximate the rank operator properly, the salient object detection result is often less precise.

3 Fast Iterative Truncated Nuclear Norm Recovery

Following the work of Shen *et al.* [12], we perform image segmentation and feature extractions to gain the features, which are then stacked vertically to form a feature matrix F . Our low-rank recovery method is then applied on F .

In the low-rank based salient object detection methods, the most critical issue is how to capture the intrinsic low-rank structure of the target properly. Conventional low-rank recovery methods apply nuclear norm heuristics to recover the low-rank structure of matrices. Nevertheless, recent work [6] have shown that nuclear norm is not a good surrogate of rank operator. Therefore, in this paper, we present to use the recent proposed *Truncated Nuclear Norm* (TNN) for a better approximation of the rank operator [6].

Given a matrix $X \in \mathbb{R}^{m \times n}$ of rank r , *Truncated Nuclear Norm* only minimizes the sum of the smallest $\min(m, n) - r$ singular values since the rank of a matrix only corresponds to the top r non-zero singular values, which can be formulated as $\|X\|_r = \sum_{i=r+1}^{\min(m,n)} \sigma_i(X)$. Based on this definition, we notice that TNN uncovers the latent low-rank structure of the target salient object as long as it exists. We then replace the nuclear norm regularization by *Truncated Nuclear Norm Regularization* to enforce the low-rank structure of the matrices in Eqn.(2):

$$(X^*, E^*) = \arg \min_{\substack{X, E \\ s.t.}} \|X\|_r + \lambda \|E\|_1 \quad (3)$$

$$F = X + E.$$

Obviously, with the specific configuration, *Truncated Nuclear Norm* is non-convex and non-smooth. Here we apply the Augmented Lagrangian Method (ALM) to solve (3) in an iterative scheme similar with the way in [10] as:

$$L(X, E, Y, \mu) = \|X\|_r + \lambda \|E\|_1 + \langle Y, F - X - E \rangle + \frac{\mu}{2} \|F - X - E\|_F^2, \quad (4)$$

where $\mu > 0$ is the penalty parameter. We can obtain the final low-rank salient object by minimizing $L(X, E, Y, \mu)$ via the following three steps:

$$X_{k+1} = \arg \min_X L(X, E_k, Y_k, \mu), \quad (5)$$

$$E_{k+1} = \arg \min_E L(X_{k+1}, E, Y_k, \mu), \quad (6)$$

$$Y_{k+1} = Y_k + \mu(F - X_{k+1} - E_{k+1}). \quad (7)$$

Before the detailed discussion on the optimization of (5) and (6), we firstly introduce a useful theorem:

Theorem 1. [2] For each $\tau \geq 0$ and $Y \in \mathbb{R}^{m \times n}$, we have

$$\mathcal{D}_\tau(W) = \arg \min_X \tau \|X\|_* + \frac{1}{2} \|X - W\|_F^2, \quad (8)$$

$$\Sigma_\tau(W) = \arg \min_X \tau \|X\|_1 + \frac{1}{2} \|X - W\|_F^2. \quad (9)$$

3.1 Optimization of Problem (5)

With fixed E_k and Y_k , the most critical issue is how to deal with the non-convex *Truncated Nuclear Norm Regularization* efficiently.

From the theoretical analysis of Hu [6], we can find a closed relationship between the *Truncated Nuclear Norm* and the traditional nuclear norm as:

$$\|X\|_r = \|X\|_* - \max_{AA^T=I, BB^T=I} \text{Tr}(AXB^T). \quad (10)$$

By adopting the same iterative scheme to relax the non-convex problem (5), in the s -th iteration, we firstly fix X_s and obtain A_s and B_s by computing the singular value decomposition of X_{k+1}^s . Then we fix A_s and B_s to update X_s . With simple algebraic operations and ignoring the constant items, we can obtain:

$$\begin{aligned} X_{k+1}^s &= \arg \min_X \|X\|_* - \text{Tr}(A_s X B_s^T) + \langle Y_k, F - X - E_k \rangle + \frac{\mu}{2} \|F - X - E_k\|_F^2 \\ &= \mathcal{D}_{\frac{1}{\mu}}(F - E_k + \frac{1}{\mu}(Y_k + A_s^T B_s)), \end{aligned}$$

where the last equality holds based on the Theorem 1. By updating the variable A, B and X alternatively, we can get the final optimal solution of (5) efficiently. The iterative scheme is summarized in Step 1 of the Algorithm 1.

3.2 Optimization of Problem (6)

With Y_k and X_{k+1} obtained from the last step, the update of E_{k+1} follows the same way in Robust Principle Component Analysis (RPCA) [10]. Through some simple algebraic operations, we can reformulate the problem (6) as follows:

Algorithm 1. Fast Iterative Truncated Nuclear Norm Recovery

Input: Observation matrix $X \in \mathbb{R}^{m \times n}, \lambda$

- 1: **Initialize** $Y_0 = \text{sgn}(D)/J(\text{sgn}(D))$, $E_0 = 0$, $X_0 = 0$ and $\mu > 0$
- 2: **for** $k = 0, 1, 2, \dots$ **do**
- 3: **STEP 1.** //Fix E , update X_{k+1} by solving problem (5).
- 4: $X_{k+1}^0 = X_k$;
- 5: **for** $s = 0, 1, 2, \dots$ **do**
- 6: $(U_s, S_s, V_s) = \text{svd}(X_{k+1}^s)$, where $U_s = (\mathbf{u}_1, \dots, \mathbf{u}_m) \in \mathbb{R}^{m \times m}$
- 7: and $V_s = (\mathbf{v}_1, \dots, \mathbf{v}_m) \in \mathbb{R}^{m \times m}$.
- 8: $A_s = (\mathbf{u}_1, \dots, \mathbf{u}_r)^T$, $B_s = (\mathbf{v}_1, \dots, \mathbf{v}_r)^T$.
- 9: $X_{k+1}^s = \mathcal{D}_{\frac{1}{\mu}}(F - E_k + \frac{1}{\mu}(Y_k + A_s^T B_s))$.
- 10: **end for**
- 11: **STEP 2.** //Fix X , update E_{k+1} by solving problem (6).
- 12: $E_{k+1} = \Sigma_{\frac{\lambda}{\mu}}(F - X_{k+1} + \frac{1}{\mu}Y_k)$.
- 13: **STEP 3.** //Update Y_{k+1} .
- 14: $Y_{k+1} = Y_k + \mu(D - X_{k+1} - E_{k+1})$.
- 15: **end for**

Output: (X^*, E^*) .

$$\begin{aligned}
 E_{k+1} &= \arg \min_E L(X_{k+1}, E, Y_k, \mu) \\
 &= \arg \min_E \lambda \|E\|_1 + \frac{\mu}{2} \|E - (F - X_{k+1} + \frac{1}{\mu}Y_k)\|_F^2 \\
 &= \Sigma_{\frac{\lambda}{\mu}}(F - X_{k+1} + \frac{1}{\mu}Y_k),
 \end{aligned}$$

where the last equality holds based on the Theorem 1.

The whole procedure of our FIT algorithm is summarized in Algorithm 1. Compared with the existing method based on RPCA [12], the main contribution of FIT is a better recovery of the low-rank structure of the background as long as it exists. Although we cannot get a closed-form solution of (5) in the iteration, the experiments show that only 10 outer and 10 inner iterations are enough to get an accurate solution. Furthermore, FIT converges faster compared with the existing salient object detection methods based on RPCA [12] (see Fig. 4(a)).

4 Experiments

In this section, we evaluate the performance of our proposed salient object detection method on the widely used 1000-image public dataset [1]. We compared our method with the state-of-the-art: ALM [12], IG [1], IT [7], LC [13], RC [4] and SR [5] from authors' implementation or results for evaluation. Among the 6 compared methods, ALM has the best performance, while our FIT outperforms ALM both on accuracy and efficiency. The experiments are conducted using Matlab on a desktop of i7 cpu and 32G memory.

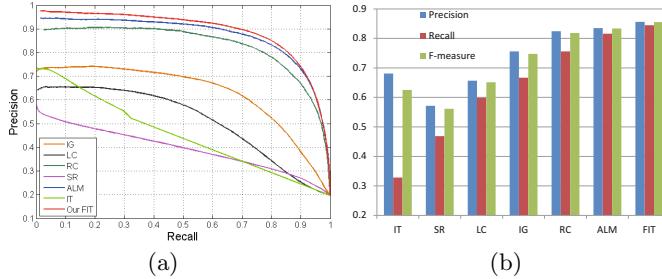


Fig. 2. Comparison result for the two evaluations based on the 7 algorithms: ALM [12], IG [1], IT [7], LC [13], RC [4], SR [5] and Our FIT. (a) Precision-recall curves on the 1000-dataset via naive thresholding. (b) Average precision, recall and F-measure on the 1000-image dataset with adaptive-thresholding segmentation. Our method outperforms the other 6 methods and thus achieves the best precision, recall and F-measure.

4.1 Evaluation Criteria

For evaluation, following the methodologies of Achanta *et al.* [1], we first apply naive thresholding to segment the saliency map according to a threshold T_f ranged in $[0, 255]$ and compare it with the ground truth mask. We then use adaptive thresholding to segment the map based on a fixed adaptive threshold T_a [1] and calculate the average precision, average recall and F-measure ($\beta=0.3$).

4.2 Feature Transformation and Higher Lever Prior Integration

Approaches have shown that human priors can further improve the detection performance. Following the work of Shen *et al.* [12], we decompose the input image and then perform a linear feature transformation learned from the MSRA dataset [11]. Higher-level human priorities (face, color and center) are then fused into the transformed image in order to further highlight the salient part.

4.3 Testing Result

A few results of the salient object detection and segmentation are shown in Fig.3. Obviously, our method generates more precise saliency maps. The backgrounds of the image in the first row are all excluded in our method while still detected some by the other methods. The center and petals of the flower in the third row are detected by our approach while mostly missing in the other methods.

For the naive thresholding, the computed Precision-Recall curves of the 7 algorithms are presented in Fig.2(a). Among these compared methods, ALM outperforms the others, and our FIT has a better performance than ALM. Furthermore, we achieve a faster convergence rate than ALM as we can see in Fig.4(a). We should note that our method could be even faster by applying GPU techniques to accelerate the multiplier operations of matrices. The computed average precision, recall and F-measure using adaptive thresholding in

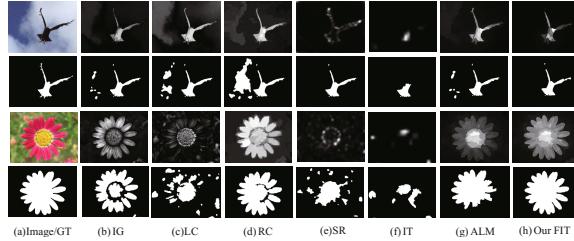


Fig. 3. Examples of extracted salient objects by different methods. The detection results of each example occupy two rows in the figure. The first column contains the original image and the correlated ground truths. The other columns represent the saliency detection result of the 7 algorithms: IG, LC, RC, SR, IT, ALM and our proposed FIT. For each example, the first row are the saliency maps, and the second row are the segmented objects using adaptive thresholding. Our FIT achieves a more accurate object segmentation among the 7 algorithms.

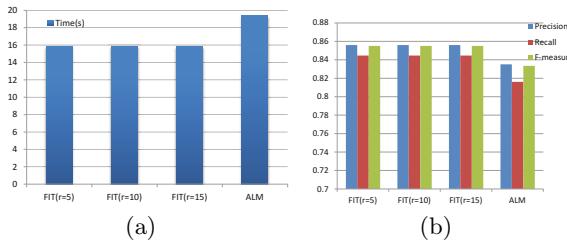


Fig. 4. Comparisons of the performance on the 1000-image dataset with r values of 5, 10, 15. (a) Time consumptions for salient object detection. (b) Average precision, recall and F-measure by adaptive-thresholding segmentation.

segmentation are shown in Fig.2(b). It's obvious that Our FIT obtains the best for each criteria, which certifies the efficiency and accuracy of our method.

4.4 Parameter Setting

For the possible values of r in $\|X\|_r$, we set the same r for all the testing images and choose the value with the best result. The comparison results for time consumption and accuracy with different r is shown in Fig.4. Since our proposed FIT is robust enough for the parameter r , we simply set r as 5 in experiments.

5 Conclusions

In this paper, we propose a novel salient object detection method called Fast Iterative Truncated Nuclear Norm Recovery. Compared with traditional nuclear norm based heuristics, FIT uses truncated nuclear norm to get a better recovery of the low-rank background. By using a simple iterative scheme, FIT can detect the salient object efficiently and accurately. Experimental results on a widely used public data set have demonstrated the effectiveness of our proposed FIT.

Acknowledgments. This work was supported by the National Basic Research Program of China (973 Program) under Grant 2013CB336500, National Natural Science Foundation of China (Grant Nos: 61125203, 61222207, 91120302).

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
2. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20(4), 1956–1982 (2010)
3. Chen, T., Cheng, M.-M., Tan, P., Shamir, A., Hu, S.-M.: Sketch2photo: internet image montage. ACM Transactions on Graphics (TOG) 28, 124 (2009)
4. Cheng, M.-M., Zhang, G.-X., Mitra, N.J., Huang, X., Hu, S.-M.: Global contrast based salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
5. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
6. Hu, Y., Zhang, D., Ye, J., Li, X., He, X.: Fast and accurate matrix completion via truncated nuclear norm regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012) (in preprint)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
8. Ko, B.C., Nam, J.-Y.: Object-of-interest image segmentation based on human attention and semantic region clustering. Journal of the Optical Society of America 23(10), 2462–2470 (2006)
9. Li, J., Levine, M.D., An, X., Xu, X., He, H.: Visual saliency based on scale-space analysis in the frequency domain. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(4), 996–1010 (2013)
10. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arxiv:1009.5055 (2010)
11. Liu, T., Sun, J., Zheng, N.-N., Tang, X., Shum, H.-Y.: Learning to detect a salient object. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
12. Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
13. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia (2006)

An Efficient Resource Allocation Method for Multimedia Cloud Computing

Yirui Li, Li Zhuo, and Haojie Shen

Beijing Signal and Information Processing Lab,
Beijing University of Technology, Beijing, China
liyirui@mails.bjut.edu.cn, zhuoli@bjut.edu.cn,
13810761814@139.com

Abstract. This paper presents a QoS-based resource allocation method for handling multimedia tasks in the cloud. Firstly, with the introduction of multiple QoS requirements of the users and the cloud service providers, the virtual resource allocation problem should be considered from three aspects: completion time, cost and energy consumption. Then the QoS-based resource allocation method is proposed. In accordance with the proposed utility function, the resource allocation results from the initial allocation will be reallocated, the best allocation results are obtained ultimately. The experimental results show that the proposed algorithm can efficiently achieve the optimal resource allocation results compared with the inherent allocation method. In addition, this algorithm also can achieve both the task utility maximization and the resource utilization maximization even under the packet loss case.

Keywords: Cloud computing, multimedia, QoS, resource allocation.

1 Introduction

Cloud computing is an emerging technology whose goal is to provide a variety of computing and storage services to multiple clients. Conceptually, cloud computing includes not only the applications and services delivered to users, but also the hardware and software in the datacenter [1]. All resources in the cloud, including architectures, platforms and software are regarded as services passed to users. From this perspective, cloud computing has advantages of high speed, high availability and high scalability.

With the development of Web 2.0, Internet multimedia is emerging as a service. To provide rich media services, multimedia computing has emerged as a noteworthy technology to generate, edit, process and search media contents, such as images, videos and so on [2]. In recent years, with the development of personal computers and mobile terminals, people are accustomed for obtaining information and services via Internet. Because of the huge amount of multimedia data and limited processing power of terminals, users usually could not obtain the service with satisfactory QoS. Therefore, it is an inevitable trend that combining cloud computing with multimedia

services. The multimedia services deployed in cloud will be effectively performed with the abundant resources and powerful processing capabilities. At the same time, users will relieve from the burden of software update and get optimal service quality.

Along with the evolution of cloud computing and multimedia cloud computing, users' requirements and application types have become more various and complex. Accordingly, there are more specific and detailed requirements for resource allocation, load balancing and scheduling control. The commercial characteristics of cloud computing and multimedia cloud computing determine that it is inevitable to introduce more factors to create the resource allocation model. Xiaoming Nan et al [3] proposed an optimal resource allocation scheme for multimedia cloud computing and simulated on the Windows Azure platform. It was one of the most forward-looking researches about this problem, but the type of multimedia tasks performed in this paper was not described clearly. By comparison, the researches about resource allocation for cloud computing had accumulated more academic results which would have use for reference to multimedia cloud computing research in turn. Reference [4] proposed a topology-aware resource allocation algorithm for data-intensive workloads. This algorithm estimated the performance of given resource allocation through prediction engine which had a lightweight emulator, then found the optimal solution by using genetic algorithm. However, this method only considered the problem of completion time, and there were not other valid parameters to characterize the performance of the system. Reference [5] proposed an agent-based resource allocation model for cloud computing. The model could adaptive allocate appropriate resource for processing users' requests based on the workload and the geographical distance between users and the datacenter. The lack of this algorithm was the limitation on type of services although it could provide fast service response time and allocation time.

In this paper, a QoS-based resource allocation method is proposed for handling multimedia tasks in the cloud. For multimedia services, not only the performance of cloud servers, but also the multiple QoS requirements should be considered while trying to solve the problem of resource allocation in the cloud. Firstly, with the introduction of multiple QoS requirements of the users and the cloud service providers, the virtual resource allocation problem should be considered from three aspects: completion time, cost and energy consumption. Then the QoS-based resource allocation method is proposed. In accordance with the proposed utility function, the resource allocation results from the initial allocation will be reallocated, the best allocation results are obtained ultimately. The experimental results show that the proposed algorithm can efficiently achieve the optimal resource allocation results compared with the inherent allocation method. In addition, this algorithm also can achieve both the task utility maximization and the resource utilization maximization even under the packet loss case.

The rest of this paper is organized as follows: Section 2 introduces the proposed resource allocation method for multimedia tasks in-depth; experimental results are presented and analyzed in Section 3; finally, the conclusion is drawn in Section 4.

2 Description of the Proposed Algorithm

Figure 1 shows the overall block diagram of the resource allocation method proposed in this paper. When users want to complete their multimedia tasks through the cloud, they will submit a request to the platform. After that, the cloud will analyze these tasks and clear the QoS requirements. Then the results will be submitted to the broker servers. The responsibility of broker servers is to compare and match those QoS requirement parameters with their own resources which are stored in the cloud. Finally, resource servers virtualize the virtual machines needed from the actual resources to perform those tasks. Ultimately, what users obtain are the final results after the execution of the tasks.

The next section will discuss the proposed method in detail.

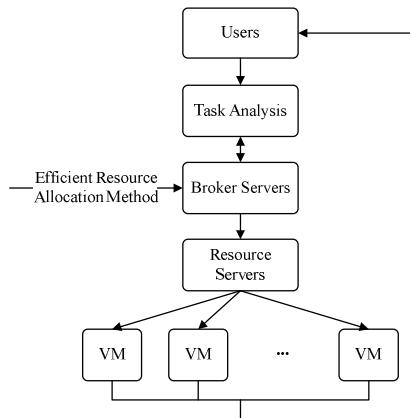


Fig. 1. An overall block diagram of the resource allocation method

2.1 Problem

With the development of Internet, the main contents of Internet services shift from static pages and small pictures to videos and other large files. The quality of multimedia application is sensitive for users. In fact, they usually could not watch videos fluently due to the bandwidth-constrained environment and the limit processing power of users' terminals. Therefore, the large-scale data processing should be moved into cloud. In this way, the workload of terminals will be reduced and the quality of services will be greatly improved.

Most complex cloud-based multimedia services are composed of multiple sub-tasks, and there are some dependent relations between them. For example, a user wants to watch a video from a website through the content delivery service which provided by the cloud. Above multimedia task includes the web page search service, the content delivery service, the broker caching service and other services. Therefore, not only the QoS guarantee of sub-tasks should be considered, but also the problem of resource competition among multiple tasks will be faced when solving the resource allocation problem.

2.2 Establishment of Utility Function

Game theory is introduced in this paper to solve the problem of limited resource competition among multimedia tasks. The design of utility function should be thought from two angles. For users, they concern more about the completion time and cost of whole multimedia tasks, because they want to spend least money for enjoying the most fluent videos. For resource providers, they concern more about the energy consumption. For these reasons, this paper redesigns the utility function which covers above two sides. The main target is to maximize users' satisfaction and resource utilization while minimizing energy consumption, ultimately achieving the enhancement of the comprehensive utility. In addition, the completion time is set as the incentive factor and the cost and energy consumption are set as the punishment factor.

Suppose n multimedia tasks share m computing resources. Each multimedia task S_i is composed of $k(i)$ sub-tasks in parallel. Each computing resource R_j has a fixed price P_j and a fixed energy consumption M_j according to its computing ability. When multiple sub-tasks are assigned to the resource, they share the computing capacity of R_j proportionally. The solution of resource allocation can be seen as a non-negative matrix a_{nm} , where each row represents a multimedia task, and each column represents a resource. In addition, a_{ij} represents the number of sub-tasks which are assigned to R_j .

According to matrix a , it is possible to obtain other three $n \times m$ matrixes, i.e. the time matrix T , cost matrix C and energy consumption matrix E . In general, there is a tradeoff between time, cost and energy consumption for each multimedia task. In this paper, ω_t , ω_c and ω_e are used to represent the weight of completion time, cost and energy consumption respectively, and $\omega_t + \omega_c + \omega_e = 1$. Formula (1) represents the utility of multimedia task S_i , the target for each task is to maximize its utility, i.e. minimize the completion time, cost and energy consumption, thereby obtaining the comprehensive improvement of users' satisfaction and resource utilization.

$$u_i(a_i) = \frac{1}{\omega_t * \max_{t_{ij} \in t_i} \{t_{ij}\} + \omega_c * \sum_{j=1}^m c_{ij} + \omega_e * \sum_{j=1}^m e_{ij}}. \quad (1)$$

The total utility of all tasks is shown in formula (2).

$$u(a) = \sum u_i(a_i). \quad (2)$$

2.3 Initial Resource Allocation

For each multimedia task, the goal is to maximize the utility value and enhance the users' satisfaction and resource utilization. As for the initial resource allocation, a multimedia task is trying to choose the most powerful resource to solve its own optimization problem without considering other multimedia tasks which share the resource with it. However, it can easily result in irrational distribution of resources by such an allocation strategy. The imbalance of resource load has led to the decline of users' QoS and reduction of the overall multimedia tasks effectiveness.

There are two cases may occur after running multimedia tasks by initial resource allocation, i.e. the resource has been reused or monopolized. If above-mentioned matrix a_{pm} satisfies the formula (3) after allocation, it indicates that the resource is monopolized, and the initial resource allocation is the unique resource optimization strategy for all multimedia tasks. On the contrary, if the formula (3) is not satisfied which means there is a resource competition problem, the operation of reallocating the resource will be done in next section.

$$\forall a_{ij} : (a_{ij} \leq 1) \cap \left(\sum_{i=1}^m a_{ij} \leq 1 \right). \quad (3)$$

2.4 Resource Reallocation

As described in Section 2.3, the reuse of resources would make users' QoS decreased, so those resources which are reused by multiple multimedia tasks should be reallocated. In this case, the results from resource reallocation must increase the effectiveness of all tasks. Once adjusting a sub-task allocation policy, all multimedia tasks will be notified with this change.

In order to evaluate the impact of the resource reallocation for multimedia tasks' utility, this paper counts the reallocation utility loss of a single task and the reallocation utility loss of global tasks respectively to measure the effectiveness of reallocation scheme. Matrix a_i means the previous resource allocation matrix of multimedia task S_i , matrix \dot{a}_i represents the resource allocation matrix when a sub-task is transformed from one virtual resource to another. The reallocation utility loss of a single task is represented by formula (4), the smaller value of this expression is, the greater improvement of the utility obtains.

$$\tau = u_i(a_i) - u_i(\dot{a}_i). \quad (4)$$

The effectiveness of a single task has been enhanced does not indicate that the result is favorable for the global tasks. When multiple multimedia sub-tasks are competing for one virtual resource, the global utility loss will decide which task has the priority of reallocation. In this paper, the multimedia sub-task which has the minimum global utility loss is chosen to reallocate in the case of multiple multimedia tasks have negative value of global utility loss. The global utility loss of reallocation is represented by formula (5), the smaller value of this expression is, the greater improvement of the global utility obtains.

$$\Gamma = \sum_{a_i \in a} u_i(a_i) - \sum_{\dot{a}_i \in a} u_i(\dot{a}_i). \quad (5)$$

The flowchart of the proposed algorithm is shown in Figure 2.

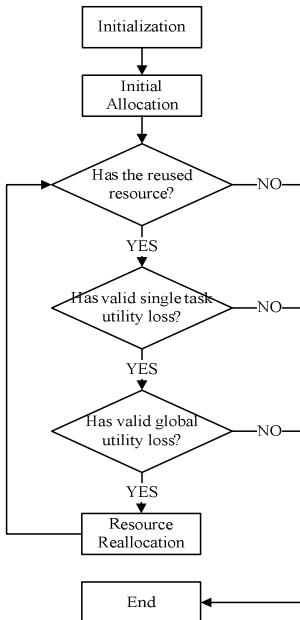


Fig. 2. The flowchart of the proposed algorithm

3 Experimental Results

In order to verify the effectiveness of the proposed resource allocation algorithm, the CloudSim is used as the simulation platform for further evaluation and analysis. The CloudSim platform has modeled cloud-based content distribution service; it sets the corresponding application complexity according to the calculation requests of the service, which ensures the validity of the simulation.

In this paper, a scene that users get the video data they want through content distribution service from the cloud has been simulated. A video sequence of 300 frames which at CIF resolution is used in this test to regard as the users' desired video data in the cloud. In the simulation, twenty virtual machines of four different types are set for users. Table 1 lists the available virtual resources' parameters in CloudSim.

Table 1. Virtual resources' parameters

Resource ID	The number of CPU	Memory (MB)	Processing speed (MIPS)
VM _{0~VM₄}	4	2048	4000
VM _{5~VM₉}	2	1024	2000
VM _{10~VM₁₄}	2	1024	1000
VM _{15~VM₁₉}	1	512	1000

The advantage of handling multimedia tasks in cloud is the cloud can perform multiple task requests through efficient parallel calculation and scheduling; therefore it can provide satisfactory QoS to users. In addition, above-mentioned advantage will be more obvious as the number of tasks increases if the proposed method has been used. Table 2 lists the comparison of relevant parameters generated from different methods when the number of tasks is 40, 60, 80, 100 and 120. The inherent method refers to the method described in Section 2.3. In particular, the results from this paper are assessed and verified through simulation environment, the results obtained in actual environment will be different because the large-scale commercial cloud platforms have their own server performance and technical indicators.

Table 2. Multimedia tasks completion time, cost and energy consumption comparison

The number of tasks	Completion time (time unit)		Average cost (cost unit)		Energy consumption (kwh)	
	Inherent method	Proposed method	Inherent method	Proposed method	Inherent method	Proposed method
40	3258.20	2968.74	853.23	761.34	3.13	2.18
60	4789.04	4447.53	996.72	872.28	3.21	2.30
80	6134.42	5495.57	1052.36	913.84	3.51	2.53
100	7450.80	6724.42	1173.04	985.77	3.62	2.64
120	8523.34	7267.30	1550.47	1236.56	3.84	2.75

With the increase of the number of tasks, Table 2 shows that the proposed method has more obvious advantages compared with the inherent method which has been mentioned before. The resource allocation method in this paper could both meet users' QoS and maximize the resource utilization while achieving load balancing. Due to network instability, packet loss usually occurs. When packet loss occurs, users usually want to get the lost packets back from the cloud, but it will result in the increase of additional completion time, cost and energy consumption. If users still could not receive the lost packets after three times retransmission, these packets will be discarded and then recovered at the clients. In this test, the packet loss rate is set as 5%, 10% and 15% respectively, and Table 3 shows the corresponding results.

From Table 3, it can be seen that although the packet loss leads to the delay of the multimedia tasks completion time, the experiment still proves that users would get more video data in a certain period of time with small amount of delay through the proposed method. By comparison, it is concluded that the delay doesn't affect the users' fluent feeling when they watch the video which they get from the cloud. Users can still get satisfactory QoS while the resources in cloud are optimal allocated through the proposed algorithm even under the packet loss case.

Table 3. Multimedia tasks completion time, cost and energy consumption comparison under the packet loss case

Packet loss rate	The number of tasks	Completion time (time unit)		Average cost (cost unit)		Energy consumption (kwh)	
		Inherent method	Proposed method	Inherent method	Proposed method	Inherent method	Proposed method
5%	40	3440.11	3134.43	891.54	798.63	3.18	2.23
	60	5029.54	4681.52	1048.52	915.75	3.32	2.35
	80	6451.75	5783.65	1102.87	959.89	3.58	2.60
	100	7833.33	7079.87	1235.21	1043.32	3.72	2.73
	120	8950.07	7641.10	1640.49	1295.86	3.92	2.80
10%	40	3590.02	3285.36	940.55	849.57	3.32	2.36
	60	5267.94	4916.97	1115.64	992.23	3.56	2.58
	80	6752.83	6071.50	1189.17	1030.20	3.79	2.78
	100	8199.87	7429.03	1297.43	1091.82	3.98	2.95
	120	9381.62	8018.37	1843.03	1446.96	4.13	3.03
15%	40	3750.83	3463.25	988.60	886.62	3.54	2.56
	60	5512.93	5156.79	1166.16	1027.64	3.63	2.64
	80	7061.60	6358.36	1211.21	1049.73	4.04	3.03
	100	8570.24	7759.40	1332.57	1116.55	4.11	3.07
	120	9805.35	8367.88	2068.08	1604.88	4.20	3.09

4 Conclusions

In this paper, a QoS-based resource allocation method for handling multimedia tasks in the cloud is proposed. A scene that users get video data through the cloud due to the limited processing power of their terminals is simulated. After that, the deployment of existing resources in cloud for providing satisfactory services to multiple users is explored in-depth. Firstly, with the introduction of multiple QoS requirements of the users and the cloud service providers, the virtual resource allocation problem should be considered from three aspects: completion time, cost and energy consumption. Then a QoS-based resource allocation method is proposed. In accordance with the proposed utility function, the resource allocation results from the initial allocation will be reallocated, and the best allocation results are obtained ultimately. The experimental results show that the proposed algorithm can efficiently achieve the optimal resource allocation compared with the inherent allocation method. In addition, this algorithm also can achieve both the task utility maximization and the resource utilization maximization even under the packet loss case.

Acknowledgments. The work in this paper is supported by the Program for New Century Excellent Talents in University (No.NCET-11-0892), the Specialized Research Fund for the Doctoral Program of Higher Education (No.20121103110017), the National Natural Science Foundation of China (No.61003289, No.61100212), and the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD201304036).

References

1. Leavitt, N.: Is Cloud Computing Really Ready for Prime Time. *Computer* 42, 15–20 (2009)
2. Zhu, W., Luo, C., Wang, J., Li, S.: Multimedia Cloud Computing. *IEEE Signal Processing* 28, 59–69 (2011)
3. Nan, X., He, Y., Guan, L.: Optimal Resource Allocation for Multimedia Cloud Based on Queueing Model. In: *IEEE International Workshop on Multimedia Signal Processing*, pp. 1–6 (2011)
4. Lee, G., Tolia, N., Ranganathan, P., Katz, R.H.: Topology-aware Resource Allocation for Data-intensive Workloads. *Computer Communication Review* 41, 120–124 (2011)
5. Jung, G., Sim, K.M.: Agent-based Adaptive Resource Allocation on the Cloud Computing Environment. In: *40th IEEE International Conference on Parallel Processing Workshops*, pp. 345–351 (2011)

Research and Application of Corrosion Prediction Based on GRA-SVR

Dongmei Fu¹, Jinlong Xiang¹, and Xiaogang Li²

¹ School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing 100083, China

² School of Materials Science and Engineering,
University of Science and Technology Beijing, Beijing 100083, China
fdm2003@163.com

Abstract. Corrosion prediction is a technology of finding the corrosion law based on material corrosion data. Due to corrosion data has the characteristics of high dimensional nonlinearity, randomness and limited sizes, many data modeling methods based on large samples are not applicable. In the process of corrosion prediction, we have to deal with missing data values, outlier detection, feature selection and regression. However, feature selection and regression would be the focus of our research in this paper. This paper adopts a modeling method combining of Grey Relational Analysis and Support Vector Regression, referred to as GRA-SVR, the former is used to select feature and the latter is used for regression. The experimental results show that, GRA-SVR method achieves higher precision than other methods such as BP Neural Network.

Keywords: Corrosion Prediction, Grey Relational Analysis, Support Vector Regression.

1 Introduction

Corrosion is a natural phenomenon that material gradually loses its function in the interaction effects of various environmental factors [1]. Corrosion prediction is to deduce long-term corrosion behavior from short term corrosion data, deduce corrosion law of large sample from local sample characteristics, and deduce corrosion behavior of the actual environment from simple indoor conditions. Corrosion prediction and corrosion test constitute the two pillars of corrosion study [2]. Compared with the corrosion test, corrosion prediction technology is still in its infancy. Corrosion prediction technology can be effectively conduct to study the corrosion behavior law of materials, estimate usage of materials and residual life, and tell people take preventive measures to avoid or reduce accidents.

Support Vector Machine (SVM) was put forward in 1995 by Vapnik and etc, which is a kind of Machine learning method based on VC dimension theory and structure risk minimization principle. Showing many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problem, makes SVM has attracted widespread attention and obtained the rapid development [3, 4]. At the same

time, the SVM application gradually expanded in the field of nonlinear regression estimation, named as SVR, and showed good performance [5]. Corrosion data is always measured in a year or longer, leading to corrosion data sets under certain conditions are usually small, and corrosion is often associated with randomness and influenced by many factors, applying SVR algorithm to material corrosion prediction is a new attempt. At the same time, the relationship between corrosion and environmental impacting factors is unclear or the relational information is incomplete. It can be difficult to feature selection with traditional statistical methods or machine learning methods. Grey Relational Analysis (GRA) is a method of Grey System Theory, which can be used to identify major correlations among factors of a system with a relatively small amount of data. Therefore, GRA can help us obtain more useful features and remove features with lower correlation. Finally, compared with the BP Neural Network (BP-NN) and Standard Support Vector Regression (Standard-SVR) without GRA, experimental results with actual corrosion data show that GRA-SVR has better performance.

2 Basic Principle

2.1 Grey Relational Analysis

In many statistical correlation analysis methods, data distribution is assumed as linear, exponential or logarithmic, and errors are normally distributed with zero means. Moreover, sufficient data are required to determine its distribution type and to ensure statistical significance [6]. However, it may be difficult to get adequate information in most corrosion prediction for the distribution of corrosion is unknown and the corrosion data of a material under specific environment is very few. Under such conditions, many traditional statistical methods based on large number of data may not be available. GRA is a quantitative description and comparison of a system development, which based on the mathematical foundations of space theory, in accordance with the four criteria as normative, integrity, symmetry and proximity, to determine relational coefficient and relation grade between the reference series and comparative series [7].

Generally, the procedure of GRA consists of the following two steps.

1. The generation of grey relation

Suppose X_i ($i=0,1,\dots,n$) is system factor, and $x_i(k)$ is the observed value of X_i at (time) point k , the behavior series of X_i was defined as follow:

$$X_i = (x_i(1), x_i(2), \dots, x_i(m))^T \quad (1)$$

The standardization of the various attributes is necessary, so that every attribute has the same amount of influence, thus the data is made dimensionless by using various techniques as initial value processing, average value processing and etc.

2. The calculation of grey relational grade

Suppose x_0 is the reference series and x_1, x_2, \dots, x_n are the comparative series. GRA uses the grey relational coefficient to describe the trend relationship between a comparative series and a reference series at a given (time) point in a system, and was defined as follow:

$$\xi_{0i}(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max_i \max_k |x_0(k) - x_i(k)|} \quad (2)$$

where ρ is a distinguishing coefficient used to control the level of differences of the relational coefficient s , and $\rho \in (0,1)$. Grey relational grade be defined as the mean of each series grey relational coefficient at all (time) points as follow:

$$\gamma_{0i} = \frac{1}{N} \sum_{k=1}^N \xi_{0i}(k) \quad (3)$$

2.2 Support Vector Regression

In recent years, due to many advantages dealing with nonlinear and small sample problem, the SVM has been widely applied in many fields, such as text classification, fault diagnosis and power load forecasting, etc. SVM algorithm was developed from the optimal hyperplane in linearly separable case, the so-called optimal hyperplane is required not only can separated this set of vectors without error and the distance between the closest vector and the hyperplane is maximal [8]. ϵ -insensitive loss function is introduced into the field of SVM regression [Vapnik,1995], abbreviated ϵ -SVR, its goal is to find a function $f(x)$ that has at most ϵ deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ϵ , but will not accept any deviation larger than that [9]. ϵ -insensitive loss function $|\xi|_\epsilon$ described by

$$|\xi|_\epsilon = \begin{cases} 0 & \text{if } |\xi| \leq \epsilon \\ |\xi| - \epsilon & \text{otherwise.} \end{cases} \quad (4)$$

The above problem can be described as a convex optimization problem as follow:

$$\begin{aligned} & \text{Minimize} \quad \frac{1}{2} \|w\|^2 \\ & \text{Subject to} \quad y_i - \langle w, x_i \rangle - b \leq \epsilon, \langle w, x_i \rangle + b - y_i \leq \epsilon \end{aligned} \quad (5)$$

In actual corrosion problems, the corrosion indicator often comes from the mean of corrosion observed value in several groups of samples for corrosion randomness, therefore, the formula (5) in the convex optimization problem is not feasible, in this case, we allow some errors to the formula (6) as shown, to deal with the constraint conditions do not satisfy the formula (5) by introducing slack variables ξ_i, ξ_i^* .

$$\begin{aligned}
& \text{Minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\
& \text{Subject to } y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i, \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*, \xi_i, \xi_i^* \geq 0
\end{aligned} \tag{6}$$

$$\begin{aligned}
L(a, w, b, \xi_i, \xi_i^*) = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n a_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\
& - \sum_{i=1}^n a_i (\varepsilon + \xi_i + y_i - \langle w, x_i \rangle - b) - \sum_{i=1}^n (\eta_i \xi_i + \eta_i^* \xi_i^*)
\end{aligned} \tag{7}$$

where L is Lagrange, and $\eta_i, \eta_i^*, a_i, a_i^*$ are Lagrange multipliers which satisfy $\eta_i^{(*)} \geq 0, a_i^{(*)} \geq 0$. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality. The optimization problem in the formula (6) can be converted into:

$$\begin{aligned}
& \text{Maximize} \sum_{i=1}^n y_i (a_i - \alpha_i^*) - \varepsilon \sum_{i=1}^n (a_i + \alpha_i^*) - \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\
& \text{Subject to } \sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in [0, C], i = 1, \dots, n
\end{aligned} \tag{8}$$

At the same time, we could obtain

$$w = \sum_{i=1}^n (a_i^* - a_i) x_i, f(x, a_i, a_i^*) = \sum_{i=1}^{n_{sv}} (a_i - a_i^*) \langle x_i, x \rangle + b \tag{9}$$

where n_{sv} is the number of support vector.

It is often difficult to describe relationship between corrosion and impact factors with linear regression model. In the case, the solution is mapping the original space into some feature space by using nonlinear transform function $x \rightarrow \Phi(x)$, and dealing with the optimization problem in the formula (8) by replacing $\langle x_i, x_j \rangle$ with $\Phi(x_i) \cdot \Phi(x_j)$. To avoid the complex operation in high dimensional space by defining Kernel function as $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ and the corresponding decision function is

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n_{sv}} (a_i - a_i^*) K(x_i, x) + b \tag{10}$$

3 Corrosion Prediction Based on GRA-SVR

In this paper, we mainly research corrosion prediction problems where the input is a vector of n components or features (environmental influential factors), therefore, corrosion data used for modeling can be written as the following matrix form:

$$(Y, X_1, \dots, X_n) = \begin{pmatrix} y(1) & x_1(1) & \dots & x_n(1) \\ y(2) & x_1(2) & \dots & x_n(2) \\ \vdots & \vdots & & \vdots \\ y(N) & x_1(N) & \dots & x_n(N) \end{pmatrix}_{m \times (n+1)} \quad (11)$$

where m is sample size, n is the number of features, Y is corrosion status indicators, $x_i(k)$ for the value of ith feature in kth sample. Corrosion prediction based on GRA-SVR mainly includes the following steps:

Data Preprocessing. In the influence of noise and other reasons, it would tend to cause missing values or outliers in the acquisition process of corrosion data. Modeling with these data directly will affect the performance of the model, so that missing data handling and outlier detection would be necessary before modeling. Simultaneously, data normalization is essential to eliminate the different dimension, otherwise, the error of modeling could be larger or even worse.

Feature Selection with GRA. Feature selection is the process of identifying and removing as much irrelevant and redundant information as possible from an original feature set for the purpose of providing better prediction accuracy [6]. In corrosion prediction problems, the sample sizes are always small, and there are many influential factors, most of feature selection methods tend to give poor results, while GRA can generate satisfactory outcomes. Corrosion status indicators is regarded as the reference series and environmental influential factors are comparative series, respectively compute the grey relational grade $\gamma = (\gamma_{01}, \gamma_{02}, \dots, \gamma_{0n})$ by using formula (2) (3), and order the features according to γ , the sorted result is $\Gamma_n = (\Gamma_{01}, \Gamma_{02}, \dots, \Gamma_{0n})$. Finally, Sequential Backward Selection (SBS) is used to the optimal feature subset search, which start with the full feature sets Γ_n , repeatedly delete the least significant feature in Γ , and end with the algorithm when the error of modeling is no longer decline with reduction of feature subset .

Regression with SVR. RBF kernel function is selected in the process of regression for it shows better performance than other kernel function, and its structure is as follows:

$$K(x, x_i) = \exp\{-\gamma \|x - x_i\|^2\} \quad (12)$$

Three parameters should be adjusted in SVR, which are insensitive coefficient ε , penalty parameter C and kernel parameter γ . There is no recognized the best way about selection of SVR parameters, the common methods include grid search algorithm, gradient descent method, genetic algorithm and so on. Grid search algorithm is a relatively simple way of parameter optimization, but it will take a much long time for its iteration times will grow quickly with the expansion of the parameter space. An improved grid search algorithm was proposed in process of parameters optimization. Compared with the traditional grid search algorithm, the difference of the improved algorithm proposed in this paper is at uniform growth with respect to parameter, which can greatly reduce the parameters optimization iterative times without affecting

the precision of the model. For example, assume a parameter p in [1,1000] and step-size equals 1, iteration times is 1000 in grid search algorithm, however, step-size is not constant in improved methods, search space of p is [1,2,...,10,20,...100,200,...,1000] and iteration times is 30. Cross Validation (CV) is a statistical analysis method used to verify model, the available subjects would be randomly split into two subsamples with one being used to fit and the other to validate [10]. In this paper, the k -CV method was used, the modeling sample was divided into k groups in this methods, each data subset would be selected as a validation set, and the remaining $k-1$ set of data as the training set, which constitutes k model, Take mean square error of those k model as the performance indicators to evaluate model. k -CV Method can effectively avoid the happening chances of over-fitting and under-fitting, and the modeling result is more convincing.

4 Experimental Results

4.1 Experimental Data

Our experimental data are taken from “The collection of The Eighth-Five corrosion data”, which is gained in “The National Natural Science Foundation Projects” of China [11]. We choose a total of 37 groups of carbon steel soil corrosion data which are collected by the material corrosion test stations located all over China for many years. The experimental data are listed below in Table 1, and Y is the Corrosion rate of steel soil, X is eleven influencing factors including buried years, PH, organic, nitrogen, HCO_3^- , Cl^- , SO_4^{2-} , Ca^{2+} , Mg^{2+} , K^+ , Na^+ .

Table 1. Partial carbon steel soil corrosion data sheet

Y	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁
1.49	1	9.9	1.25	0.059	0.058	0.009	0.013	0.006	0	0.001	0.063
1.27	3	10	0.8	0.046	0.0757	0.0038	0.0154	0.0054	0.0008	0.0055	0.0402
3.61	4	5.5	0.28	0.01	0	0	0.0061	0.0007	0.0007	0	0.0001
2.94	7	5.4	0.37	0.018	0.0031	0.0014	0.0036	0.0014	0.0004	0.0002	0.0012
3.22	3	6.9	2.21	0.113	0.0168	0.0008	0.0097	0.0037	0.0011	0.0002	0.0046
4.16	1	8	0.21	0.033	0.0205	0.0007	0.0133	0.0064	0.0021	0.0003	0.0029
2.34	3	8.1	0.26	0.027	0.0168	0.0017	0.0114	0.0059	0.0009	0.0007	0.0029
3.95	3	8.4	0.54	0.026	0.0207	1.0945	0.112	0.0197	0.0384	0.0249	0.6239
2.84	4	4.9	0.73	0.052	0.0008	0.0007	0.0074	0.0029	0.0011	0	0
2.16	6	4.6	0.47	0.046	0.0028	0.0043	0.0004	0.0022	0.0005	0.0001	0.0002
.....											

4.2 Results and Analysis

Mean square error (MSE) and the mean relative error (MRE) are two commonly used error criterion to measure the performance of the model, which are defined as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (observed_i - predicted_i)^2, MRE = \frac{1}{n} \sum_{i=1}^n \frac{|observed_i - predicted_i|}{observed_i} \cdot 100 \quad (13)$$

Four experimental data samples were selected randomly as test data and the others serve as training data. The SVR algorithm used in this paper was implemented based on LIBSVM [12]. SVR parameters satisfy $\varepsilon \in [0.01, 0.1]$, $C \in [1, 1000]$, $\gamma \in [0.01, 10]$, and CV fold k is 10. GRA distinguishing coefficient ρ is 0.5. In experimental process, the training data were applied to modeling with BP-NN, Standard-SVR and GRA-SVR respectively, and the performance of the above three models respectively with the test data were verified, finally, the results were shown in Fig.1 and table 2.

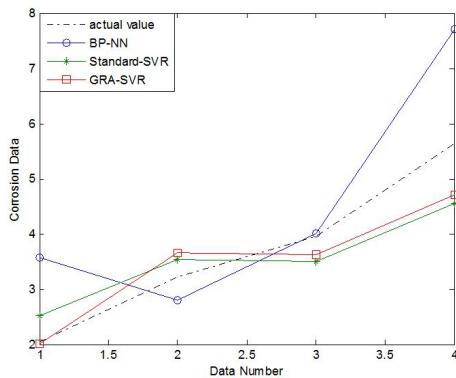


Fig. 1. The forecasting performance of the three models

Table 2. Forecasting Error of the three models

Error	BP-NN	Standard-SVR	GRA-SVR
MSE	1.69	0.43	0.30
MRE (%)	31.25	16.08	10.10

The results show that the precision of two kinds of SVR regression methods were better than the BP neural network, as support vector machine is proper to deal with small sample data naturally because of sparsity; Moreover, GRA-SVR has a better generality than the standard-SVR without GRA, for when we consider the GRA, we emphasize those attributes who have heavy relevancy to the output, to eliminate ones who have little relevancy that may cause perturbations in model. And using the grey method to compute relevancy instead of other methods which are based on possibility, it represents good adaptability to small sample data. The advantages of support vector and GRA hopefully could bring us a reasonable model, like GRA-SVR.

5 Conclusions

The characteristics of the material corrosion data were systematically summarized in this paper, GRA-SVR method combining the strengths of GRA and SVR was applied

to material corrosion prediction. To ensure the generalization ability and execution efficiency, an improved grid search algorithm based on cross validation was used during regression. Experimental results show that the performance of both GRA-SVR and Standard-SVR are better than BP-NN, which reflects the advantage of the SVR dealing with nonlinear data with small sample, meanwhile, GRA could give favor to SVR, it is obvious from higher accuracy than Standard-SVR. Applying GRA-SVR to material corrosion prediction is a new attempt, which has great reference significance to practical problems such as material choice and design.

References

1. Li, X., Guo, X., He, Y.: Corrosion and Prediction of Materials. Central south university press, Hunan (2009)
2. Weng, Y.: Corrosion Prediction and Basic Chemometrics. Journal of Chinese Society for Corrosion and Protection 31(8), 245–249 (2011)
3. Vapnik, V.N.: An Overview of Statistical Learning Theory. IEEE Transactions on Neural Networks 10(5), 988–999 (1999)
4. Gu, Y.-X., Ding, S.-F.: Advances of Support Vector Machines(SVM). Computer Science 38(2), 14–17 (2011)
5. Vapnik, V.N., Golowich, S.E., Smola, A.J.: Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing, Systems 9, 281–287 (1996)
6. Song, Q., Shepperd, M.: Predicting software project effort: A grey relational analysis based method. Expert Systems with Applications 38(6), 7302–7316 (2011)
7. Liu, S., Dang, Y., Fang, Z.: The grey system theory and its application. Science press, Beijing (2010)
8. Zhang, X.: Introduction to Statistical Learning Theory and Support Vector Machines. Acta Automatica Sinica 26(1), 32–41 (2000)
9. Smola, A.J., Schlkopf, B.: A tutorial on Support Vector Regression. Statistics and Computing 14(3), 199–222 (2004)
10. Hawkins Douglas, M., Basak Subhash, C., Denise, M.: Assessing model fit by cross-validation. Journal of Chemical Information and Computer Sciences 43(2), 579–586 (2003)
11. Hou, X., Cao, C.: The collection of the Eighth-Five Corrosion data (the collection of the soil erosion data). Institute of Metal Research Chinese Academy of Sciences 3 (June 1996) project No.59290900-03-03
12. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 27(2), 1–27 (2011)

Research on a Super-Sparse Data Generation Model for Temperature Data Map

Dongmei Fu¹, Wuchen Li¹, Xingen Li², and Xiaoming Wang²

¹ School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing 100083, China

² Shandong Electric Power Research Institute,
Jinan 250000, China
fdm2003@163.com

Abstract. Data visualization is an important method in data mining, and temperature data map is one of the methods of data visualization. When sample temperature points are limited, it brings the main content of the research on how to build a temperature data map, namely how to use super-sparse date to generate a data map. Based on the fact that temperature energy keeps constant, this paper proposes a diffusion model which uses the known temperature data to predict the unknown values. Furthermore, this paper improves the model by adding the influence of wind to make a better temperature data map. Taking the area of Shandong Province as an example, we can only get 17 cities' annual average temperature, then how to make a whole map of Shandong area confronts us. Experimental results verified that the proposed diffusion models turn out to be feasible.

Keywords: diffusion model, data map, wind influence.

1 Introduction

Data visualization is a study of the data's visual representation on the screen using computer graphics, image processing technology, and the technology of human-machine interaction. It turns data into a geometric representation, and allows people to observe the desired simulation and calculation results[1-5]. Data map is one of the most intuitive and simplest data visualization methods. Each point in the data map should stand for a data value[6,7]. Current methods always make an icon on behalf of the statistical data in the map based on statistics, or use a single color value to represent the whole region's statistical data. However, the data map generated with this method doesn't match data's real distribution on the space, and in this kind of map the difference within the same level is ignored, namely the gradual change in the data is replaced with saltation on the boundaries[8].

In this paper, we study the method of data map generation in a gradually transitional way. The map with gradually changing data has been widely used. For example, to evaluate the environmental influence on material corrosion, a certain number of corrosion test sites have been set up worldwide. These sites constantly monitor the

material corrosion rate and environmental parameters such as temperature, wetness, chloride ion concentration etc which are related to the material corrosion rate closely. To apply the experiment results from these sites to other regions, the environmental parameters should be first known there. Then the map with gradually changing data becomes a good choice which is an intuitive and desirable method. However, limited by the number of experiment sites, very little data can be collected. Especially the air wetness, chloride ion concentration, SO_2 concentration etc related closely to corrosion[9, 10] are particularly rare. Therefore, it's very necessary to study a method of data map generation with the super-sparse data.

2 Theory of Temperature Diffusion Model

The notation is defined as follows: P represents the area expected generation temperature data map, T_{Avg} is the average temperature of P . And then we suppose the air media is homogeneous, and remove the influence of wind and altitude. According to the thermal diffusion theory, we propose the method to generate temperature data map as follows: the temperature of point A in area P is T_A , the energy diffusion function on the line l which is through the point A in Fig.1.a is

$$T = T_{\text{Avg}} + f(d) \quad (1)$$

Where d represents the distance from any other point to A, T represents temperature. When $T_A > T_{\text{Avg}}$, point A will spread energy to other places, and the diffusion of energy decreases as the distance increases, so $f(d) > 0$. Otherwise, $f(d) < 0$.

Definition 1. According to (1), we can calculate the value of any point B in area P, named as $T(A, B)$, then we definite $T(A, B)$ is point A's generation value on point B. There are m points in P , defined as $Q = \{B_i, i=1, 2, \dots, m\}$, then we define $Q_A = \{T(A, B), \forall B \in Q\}$ as point A's generation domain in area P. If there're n known points in P , then

$$Q = \sum_{i=A}^n w_i Q_i \quad (2)$$

Where Q is n known points' composite domain, w_A is point A's weight in area P.

3 The Ideal Model of Temperature Data Map

3.1 Diffusion Function of Each Data Point

In area P, there're n known temperature values, T_{Avg} is the average of them. The coordinate of point A is (x_a, y_a) , and B is (x_b, y_b) . Then the distance between them is $d(A, B) = d(B, A) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$.

When $A=B$, $d(A, B) = 0$, and $d_{A_{\max}} \geq d(A, B)$, B stands for any point in area P .

According to the temperature diffusion model, we can get the equation (3).

$$T_B = T_{\text{Avg}} + b * e^{-a^2 d(A, B)} \quad a > 0, b \neq 0 \quad (3)$$

The model should satisfy the following two conditions:

- ① If $d(A, B) = 0$, then $T_A = T_B$;
- ② If $d(A, B) = d_{A_{\max}}$, then $T_A = T_{\text{Avg}} + \gamma$.

Where $\gamma > 0$, and it should be small enough compared to T_{Avg} . In addition, $\ln(\gamma)$ also should be small. So the range of γ is $[0.5, 2]$.

According to ①②, we get the solution

$$b = T_A - T_{\text{Avg}}, a = \frac{\ln(|b|) - \ln(\gamma)}{d_{A_{\max}}} \quad (4)$$

3.2 Calculate the Generation Domain and Generate Data Map

There're n known temperature data points, then based on equation (3) we can calculate every point's generation domain $Q_i, i=1, 2, \dots, n$. Besides, the composite value on the point A of the n known points should obviously be equal to T_A , so we get the equation as follows:

$$w_A * T(A, A) + w_B * T(B, A) + \dots + w_n * T(n, A) = T_A \quad (5)$$

And we have n known points, so we can get the matrix equation

$$T * W = Y, \text{ where } T = \begin{bmatrix} T(A, A) & T(B, A) & \dots & T(n, A) \\ T(A, B) & T(B, B) & \dots & T(n, B) \\ \vdots & \vdots & \ddots & \vdots \\ T(A, n) & T(B, n) & \dots & T(n, n) \end{bmatrix}, W = \begin{bmatrix} w_A \\ w_B \\ \vdots \\ w_n \end{bmatrix}, Y = \begin{bmatrix} T_A \\ T_B \\ \vdots \\ T_n \end{bmatrix} \quad (6)$$

Matrix W can be solved, and then we can calculate the data map according to equation (2).

4 The Temperature Diffusion Model with Influence of Wind

In actual state energy diffusion is effected by the influence of wind, altitude, air density etc. The influence of wind is the most important, so we must add it to the model.

With the influence of wind, diffusion of energy will change as the wind speed and wind direction change. If $T_A > T_{\text{Avg}}$, the downwind place of energy source will get more energy, and the upwind place gets less, so the temperature of downwind place has to be higher than the upwind place. On the contrary, if $T_A < T_{\text{Avg}}$, the temperature of downwind place has to be lower than the upwind place.

Here the range of energy diffusion with influence of wind is approximated to be the shape of ellipse, and the source point A of energy diffusion should consistently stay on the focus F_1 , the other focus F_2 is on the extension of wind direction shown as Fig.1.b. If there is no influence of wind, the diffusion range of point A is a circle like Fig.1.a. If influence of wind exists, the energy diffusion range of point A is an ellipse, so the diffusion function of point A will change in every direction. In the ellipse, the line L_1 stays exactly in upwind direction of point A, L_2 in downwind direction. So L_1 gets the least energy and L_2 gets the most.

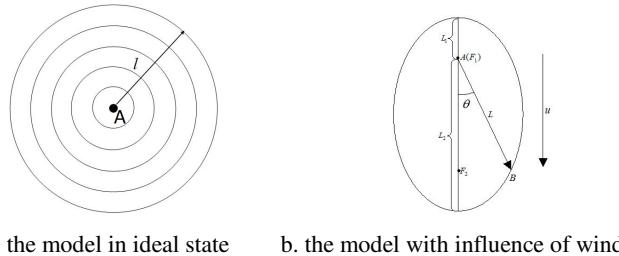


Fig. 1. Schematic diagram of the energy diffusion

How to build up the diffusion model is following:

1) Ellipse model construction

At first, let wind speed u and ellipse's eccentricity ε have the function relationship $\varepsilon = g(u)$. The equation of ellipse is $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

Let circle's area equal to ellipse's, here we define $R = d_{A\max}$, and According to the property of ellipse, we get the three parameters a, b, c through (7).

$$\pi R^2 = \pi ab, \quad \varepsilon = \frac{c}{a}, \quad a^2 = b^2 + c^2 \quad (7)$$

In Fig.1.b, point A stay on focus F_1 , and the distance between A and B is $d(A,B)$, noted as L . The angle between \overline{AB} and u is $\theta(A,B)$, then we get equation:

$$L = \frac{a(1-\varepsilon^2)}{1-\varepsilon \cos \theta} \quad (8)$$

2) Diffusion function calculation

Similar with the diffusion function in ideal state, the diffusion function with influence of wind is

$$T_B = c + b * e^{-a * d(A,B)} \quad (9)$$

The model satisfies the following three conditions:

- ① If $d(A,B) = 0$, then $T_A = T_B$, so $T_A = c + b$

② If $d(A, B) = R$, then $T_B = T_A - \Delta T$, so $T_A - \Delta T = c + b * e^{-a^* R}$

Where ΔT is the temperature variation from T_A when $d(A, B) = R$, and it can be calculated by equation $\Delta T = (T_A - T_{Avg}) * e^{1-\frac{L}{R}}$. It guarantees that when $L = R$, ΔT is the same as that in ideal state; when $L > R$, the direction of L distributes more energy and ΔT is smaller; when $L < R$, the direction of L distributes less energy and ΔT is bigger.

③ With or without the influence of wind, the whole diffusion energy from point A should be constant. Here we assume the total energy of each direction through point A keeps constant. In Fig.1.a, the total energy on any diameter of the circle is

$$\Phi = 2 \int_0^R T_{Avg} + b' e^{-a' x} dx \quad (10)$$

Where a', b' refer to the a, b in section 3. As shown in fig 2, the length of two segments through focus F_1 is L_1, L_2 . Then the total energy on the segment L_1 is

$$\Phi * \frac{L_1}{L_1 + L_2} = \int_0^{R_1} c + b e^{-ax} dx \quad (11)$$

Through equations above, the parameters in equation (9) can be solved, then the final temperature data map can be obtained through equation (2)(5)(6).

5 Simulation

Here're 17 cities' annual average temperature values, from which the temperature data map in Shandong is asked to make. According to the description in Section 4, the improved diffusion model with wind speed $u = 0$ is the same as the model in ideal state. So we only simulate the improved model, assuming that wind is from north.

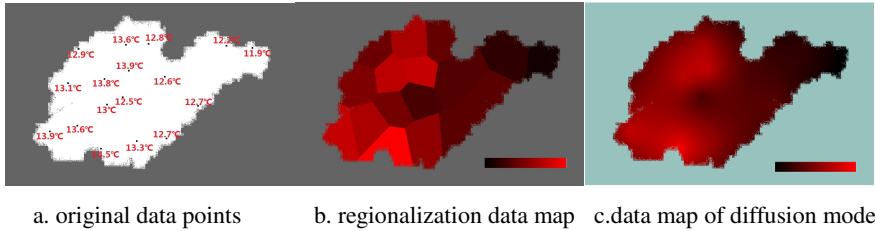
Table 1. Annual average temperature of cities in Shandong(°C)

Address	Heze	Jining	Zaozhuang	Linyi	Tai'an	Liaocheng	Laiwu	Jinan
Temperature	13.9	13.6	14.5	13.3	13	13.1	12.5	13.8
Dezhou	Zibo	Binzhou	Weifang	Rizhao	Dongying	Qingdao	Yantai	Weihai
12.9	13.9	13.6	12.6	12.7	12.8	12.7	12.2	11.9

Firstly, the data in table 1 should be mapped to [10,255], then the influence of north wind is added to the model. The wind u (m/s) and ellipse's eccentricity ε have the relationship: $\varepsilon = u / 30, u < 30$.

1) Simulation of regionalization data map

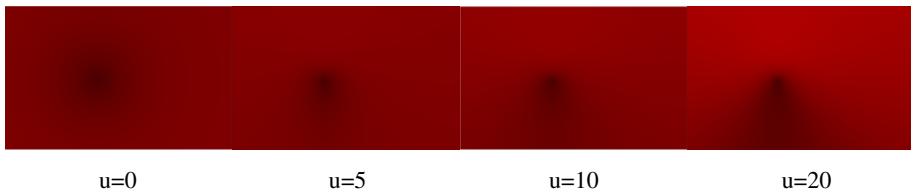
The method of regionalization data map is: the value of unknown point B in the map is equal to the value of the known point which is closest to B[11,12]. Fig.2.a is the map with known points value. Fig.2.b shows the regionalization data map, and Fig.2.c presents the result of diffusion model method.

**Fig. 2.** Different data maps

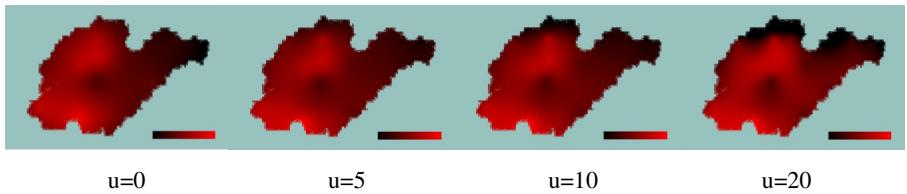
According to Fig.2, we figure out that the characteristic of gradually changing of data in the regionalization data map is totally ignored, and the method of diffusion model, which overcomes this shortcoming, shows the characteristic of data well.

2) Simulation of Diffusion model

Laiwu, the known point, is a low-temperature point and we add the influence of north wind to it. The wind speed respectively takes 0, 5, 10, 20. Then the generation domain of Laiwu is shown in Fig.3, where the black point is Laiwu.

**Fig. 3.** Generation domain of Laiwu with influence of north wind

Then we add influence of wind to every known points, the results are shown in Fig.4.

**Fig. 4.** Composite domain of all known points with influence of north wind

As shown in Fig.4, we successfully make a temperature data map filled with hundreds of thousands of pixels through the known 17 points and achieve the reasonable transition between different data points. Besides, we can get any point's temperature value t by its color value T . With the influence of north wind on point Laiwu, Fig.3 shows that the temperature in downwind place of low-temperature point becomes lower as wind speed increases. It implies the energy concentrate in upwind place. On the contrary, we can predict that with the influence of north wind, the temperature in downwind place of high-temperature points becomes higher as wind speed increases.

It implies the energy concentrates in downwind place. In Fig.4, it shows with the influence of north wind on every known point, the whole energy moves southward, and it becomes more obvious as wind speed increases.

3) Prediction with the model

Let 16 points in table 1 be known points, the other one be unknown point to be predicted. Through temperature diffusion model in ideal state and regionalization data map separately, the absolute errors of two methods are shown in Fig.5. And the cities of Jining, Linyi, Tai'an, Jinan, Weifang are in the middle of the map, surrounded by the other cities.

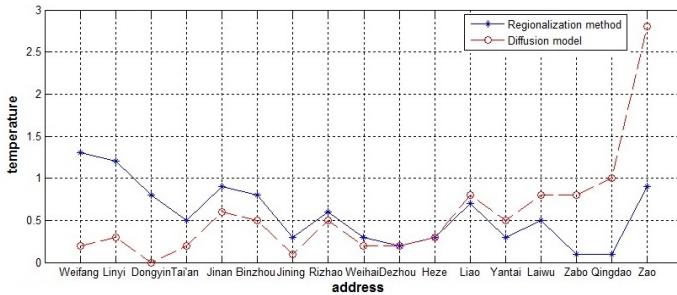


Fig. 5. Errors predicted with two methods

Among all the errors with the method of diffusion model in Fig.5, Zaozhuang's and Qingdao's are bigger, because those two cities locate on the edge of the map. The other 15 cities' errors are within 1°C, especially the 5 cities above in the middle place whose errors are within 0.5°C. It implies that the diffusion model in ideal state can well predict the temperature of the points surrounded by the known points. But points on the edge still get a worse error. Compared to regionalization data map, the diffusion model gets a better result predicted. Fig.5 presents 11 cities' errors, which contains the 5 cities' in the middle place, with method of diffusion model are not bigger than regionalization data map.

6 Conclusion

This paper simulates the diffusion model, analyses and compares the simulation results. It turns out the method of data map with diffusion model is feasible, and it presents the data in a gradually changing way successfully. This model method has a good extension. It's easy to add the influence to the model, for example this paper adds the influence of wind successfully. Besides, the method of diffusion model is of great significance. Through the method, we can just need a few sites' temperature to make the whole area's temperature data map, i.e. the method can estimate the unknown place's temperature. What's more, diffusion model method can be applied to other measurement factors with the characteristic of diffusion, like air wetness, Cl^- , SO_2 concentration etc which are closely relative to corrosion. In this way the method solves the problem that people cannot evaluate the corrosion level of places where there're no monitoring sites or lack of corrosion data.

And this method is a preliminary exploration on a super-sparse data generation model for data map. So the method needs further improving. Although the influence of wind has been added to the improved diffusion model, the consideration of other factors like terrain, air media etc, which also have effects on the diffusion, have not been considered.

References

1. Keim, D.A.: Visualization techniques for mining large databases: A comparision. *IEEE Transactions on Knowledge and Data Engineering* 8(6) (1996)
2. Li, J.X.Z.: Visualization of high-dimensional data with relational perspective map. *Information Visualization* 3, 49–59 (2004)
3. Fang, L., Kai, T.: Visualization of financial data based on SOM and gravitational field clustering. *Journal of Computer Aided Design and Computer Graphics* 24(4), 435–442 (2012)
4. de Oliveira, M.C.F., Levkowitz, H.: From visualization to visual data mining : A survey. *IEEE Trans. on Visualization and Computer Graphics* 9(3), 378–394 (2003)
5. Sun, Y., Tang, J., Tang, D.: Improved Multivariate Data Visualization Method. *Journal of Software* 21(6), 1462–1472 (2010)
6. Guo, L., Li, L., He, Z.: Visual Search Processes and Cognitive Efficiency of Statistical Maps. *Geomatics and Information Science of Wuhan University* 27(6), 637–641 (2002)
7. Yong, W., Zhinong, Z.: Research on visualization of massive graph data. *Application Research of Computers* 29(9), 3216–3220 (2012)
8. Kansheng, Y.: *The curriculum of modern cartography*. The Science Press (2009)
9. Hao, X., Li, X., Dong, C.: Grey relational analysis of atmospheric corrosion of stainless steels in terms of exposure time in representative areas. *Journal of University of Science and Technology Beijing* 30(5), 505–508 (2008)
10. ISO 9223-1992(E) Corrosion of metals, and alloys – corrosivity of atmospheres – classification
11. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Publishing House of Electronics Industry (2012)
12. Shi, P., Rui, X.: Comparison and improvement of spatial rainfall interpolation methods. *Journal of Hohai University (Natural Sciences)* 33(4), 361–365 (2005)

Strip Flatness and Gauge Multivariable Control at Cold Tandem Mill Based on Fuzzy RBF Neural Network

Li Wang

Key Laboratory for Advanced Control of Iron and Steel Process (Ministry of Education)
School of Automation & Electronic Engineering
University of Science and Technology Beijing
Beijing, 100083, China
w13927@126.com

Abstract. In the process of strip rolling, flatness and gauge control system is a time-delay, coupled and nonlinear complex system. This paper applies fuzzy RBF neural network (FRBF) to cold tandem rolling, and presents a kind of strip flatness and gauge multivariable adaptive control system. The simulation results show that this kind of new controller has good performances of adaptively tracking target and resisting disturbances and is superior to the conventional decoupled PID control in improving the strip flatness and gauge accuracy.

Keywords: neural network control, flatness control, gauge control, RBF neural network.

List of Symbols

- S: roll gap
P:rolling force
F: work roll bending force
 C_p :elastic constant of stand
Q: plastic constant of strip
 C_F :work roll bending coefficient
 ΔX : the fractional change of the variable x
 H_c, H_e :center gauge and edge gauge of entry strip
 h_c, h_e : center gauge and edge gauge of exit strip
 L_c, L_e :center length and edge length of entry strip
 l_c, l_e : center length and edge length of exit strip
 K_p :mill transverse stiffness coefficient
 K_F : transverse bending roll stiffness coefficient
 ω :complex roll mould

1 Introduction

Strip is one of major industrial products. Its important quality index is flatness and gauge. The accuracy of final strip flatness and gauge in the cold tandem mill depends

on the automatic flatness control (AFC) and automatic gauge control (AGC) at each stand, and it is well known that the strip control performance of the first stand has great influence on the final accuracy of the cold strip products. However, there are problems of interference between the strip flatness and gauge in the cold tandem mill, So it is a important problem that how to complement the AFC-AGC complex control[1]. It is also difficult to control both strip flatness and gauge with the high accuracy by means of the conventional PID (proportional integral derivative) control law. In this paper, the complex multivariable control problem of the first stand is studied by a kind of fuzzy RBF neural network control method based on analyzing strip rolling process.

First, the strip rolling process is analyzed by the method that the AFC and AGC are taken as a whole process by obtaining a complex AFC-AGC system.

Second, because AFC-AGC is a multivariable, coupled and nonlinear system, the FRBF neural network method is firstly adopted to construct a AFC-AGC multivariable complex control system. In order to fast the learning speed of RBF neural network, a new weights training algorithm is applied, meanwhile a effective center selection method of RBF network is used.

Third, simulation is done based on data collected from an Iron and Steel factory. The simulation results show that this kind of new controller has good performances of adaptively tracking target and resisting disturbances and is superior to the conventional decoupled PID control in improving the strip flatness and gauge accuracy.

Finally, the conclusion is drawn, namely, the control method in AFC-AGC is efficient.

2 A Complex System Model of Flatness and Gauge

In the first stand it is given that there are flatness control measure of work roll bending and gauge control measure of hydraulic reduction actuator, furthermore work roll bending system and hydraulic actuator system are similar to the following transfer function:

$$G_1(S) = \frac{K_1}{1 + T_1 S} \quad (1)$$

$$G_2(S) = \frac{K_2}{1 + T_2 S} \quad (2)$$

In order to going on simulation research, firstly sets up mathematic model of strip flatness and gauge complex system.

a. Linearization of gauge equation (generalized elastic deformation and plastic deformation equation) around the operation points

$$\Delta h = \Delta s + \frac{\Delta P}{C_P} + \frac{\Delta F}{C_F} \quad (3)$$

b. Flatness equation

Flatness equation can be obtained from following equations:

① Flatness equation[2]

$$\frac{I_c - I_e}{I} = \frac{L_c - L_e}{L} + \frac{H_c - H_e}{H} - \frac{h_c - h_e}{h} \quad (4)$$

② Transverse tension deviation of entry and exit strip

$$\begin{cases} \sigma_0 = -E \left(\frac{L_d}{L} \right) \\ \sigma_1 = -E \left(\frac{I_d}{I} \right) \end{cases} \quad (5)$$

③ Transverse gauge deviation equation of exit strip

$$h_d = \frac{P}{K_p} - \frac{F}{K_F} + \omega \quad (6)$$

flatness equation can be obtained from equation ①, ② and ③, namely

$$\Delta\sigma_1 = \frac{E}{h} \left(\frac{\Delta P}{K_p} - \frac{\Delta F}{K_F} - \frac{h}{H} \Delta H_d + \frac{h}{E} \Delta\sigma_0 \right) \quad (7)$$

c. Rolling force equation

plastic deformation equation of strip is as follows:

$$h = H - (P / Q) \quad (8)$$

According to equation (3) and (8), linearization of rolling force equation is deduced around the operation point, namely

$$\Delta P = \frac{C_p Q}{C_p + Q} \left(\Delta H - \Delta S - \frac{\Delta F}{C_F} \right) \quad (9)$$

It is given that all kinds of detector elements are approximately considered as first-order inertia, then flatness and gauge complex system model (equation 10) (shown in Fig.1) is available through equation (3), (7) and (9).

$$\begin{cases} \Delta h = \frac{C_p}{C_p + Q} \Delta S + \frac{Q}{C_p + Q} \Delta H + \frac{C_p}{C_p + Q} * \frac{1}{C_F} \Delta F \\ \Delta\sigma_1 = \frac{E}{h} \left[\frac{1}{K_p} * \frac{C_p}{C_p + Q} (\Delta H - \Delta S) - \left(\frac{1}{K_p} * \frac{C_p}{C_p + Q} \right. \right. \\ \left. \left. * \frac{1}{C_F} + \frac{1}{K_F} \right] \Delta F - \frac{h}{H} \Delta H_d + \frac{h}{E} \Delta\sigma \right] \end{cases} \quad (10)$$

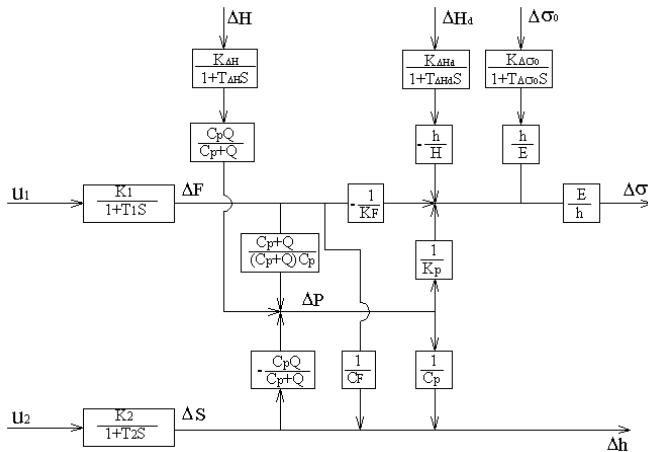


Fig. 1. Flatness and Gauge Complex System Model

3 Strip Flatness and Gauge Multivariable Control System Based on FRBF Neural Network

3.1 FRBF Neural Network Multivariable Control System

It has been proved that function equivalent relationship exists between fuzzy system and RBF (radial basis function) network under some restrict conditions, consequently it establishes theoretic foundation for RBF application in fuzzy system. In addition, RBF network can be substituted for any arbitrary input/output relationship and its structure parameters can carry out learning separately. Due to these characteristics, RBF network can be used in controlling complex and nonlinear system effectively. As a result, RBF network and fuzzy theory have emerged to constitute a self-adaptive fuzzy control system[3], which can computes quickly and has good control performance and self-adaptive ability.

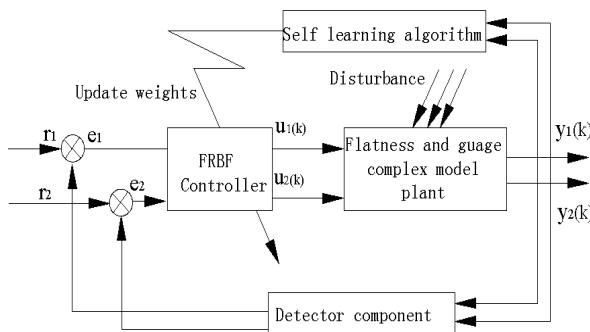


Fig. 2. Flatness and gauge FRBF control system

On account of flatness and gauge MIMO system, applies fuzzy RBF network multivariable control system to design following system whose topology structure is shown Fig.2.

where r_1 is flatness set value, $y_1(k)$ is flatness output value, r_2 is gauge set value, $y_2(k)$ is gauge output value, $u_1(k)$ is flatness regulator value--work roll bending force, $u_2(k)$ is gauge regulator value--roll gap, e_1 is error between flatness set value and output value and e_2 is error between gauge set value and output value.

In this control system, RBF network is a three layers forward network including a hidden layer, and hidden layer output function is Gaussian function. The i th neural output $R_i(r)$ in hidden layer is as follows:

$$R_i(r) = \Phi(\|r - c_i\|) = \exp\left[-\frac{(r - c_i)^2}{2\sigma_i^2}\right]$$

here r is input vector, c_i and σ_i represents the central point and width of radial basis function, respectively, $\Phi(\bullet)$ is basis function, and $\|\bullet\|$ is Euclid norm.

RBF network output is as follows:

$$y_i = \sum_{j=1}^m w_{ij} R_j(r)$$

here m is nodes of hidden layer, y_i represents the i th output, and w_{ij} is connected weight from the j th hidden layer node to the i th output layer node.

3.2 Flatness and Gauge Multivariable Controller Algorithm Based on FRBF

In this flatness and gauge multivariable control system, controller design is based on FRBF, and its design parameters include mainly center points c_i , width coefficient σ_i and weights w_{ij} . In order to meet the demand of real-time control and accomplish learning quickly, this FRBF controller adopts separate learning method of network parameters, namely, center points c_i , width coefficient σ_i and w_{ij} learn separately.

(1) Selection Method of RBF Network Center Points

It is most difficult to select center points[4] in application of RBF network. The center points of Flatness and gauge multivariable controller can obtain from large numbers of site data by making the use of recurrence and clustering algorithm[5]. Detailed steps are as follows:

- ① . According to fuzzy subset of input/output data set up center points initialization $c_i(0)$, $1 \leq i \leq m$ and learning speed η ($0 < \eta < 1$).
- ② . Calculate distance between center points and input vector, and search node with minimization

$$\mathbf{d}_i(k) = \|\mathbf{r}(k) - \mathbf{c}_i(k-1)\|, \quad 1 \leq i \leq m,$$

$$\mathbf{d}_{\min}(k) = \min_i \mathbf{d}_i(k) = \mathbf{d}_r(k)$$

③. Update center points

$$\begin{aligned}\mathbf{c}_i(k) &= \mathbf{c}_i(k-1), 1 \leq i \leq m, i \neq r \\ \mathbf{c}_r(k) &= \mathbf{c}_r(k-1) + \eta(\mathbf{r}(k) - \mathbf{c}_r(k-1))\end{aligned}$$

④. Calculate the distance of node r

$$\mathbf{d}_r(k) = \|\mathbf{r}(k) - \mathbf{c}_r(k)\|$$

Because learning rules is linear, it is sure that error convergence speed is quick.

(2) Weights online self-learning algorithm of FRBF based on process optimization

In FRBF controller shown in Fig.2, in the light of steepest descent method, weight

$w_{ij}(k)$ update algorithm is as follows:

$$w_{ij}(k+1) = w_{ij}(k) - \beta \frac{\partial E}{\partial w_{ij}(k)}$$

where $\beta (0 \leq \beta \leq 1)$ is learning coefficient, E is performance index.

$$E(k) = \sum_{t=1}^2 [r_t(k) - y_t(k)]^2$$

Here

$$\frac{\partial E(k)}{\partial w_{ij}(k)} = - \sum_{t=1}^2 e_t(k) \frac{\partial y_t(k)}{\partial u_j(k)} \frac{R_j}{\sum_{i=1}^m R_i}$$

So, weights online self-learning algorithm of FRBF based on process optimization is as follows:

$$w_{ij}(k+1) = w_{ij}(k) + \gamma \cdot \beta \cdot \sum_{t=1}^2 e_t(k) \cdot \text{sgn} \frac{\partial y_t(k)}{\partial u_j(k)} \left| \frac{\Delta y_t(k)}{\Delta u_j(k)} \right| \cdot \frac{R_j}{\sum_{i=1}^m R_i}$$

where $\gamma (0 \leq \gamma \leq 1)$ is expert estimation coefficient.

4 Simulation and Analysis

System structure is shown Fig.2.

It is given that error e_1, e_2 have respectively 7 fuzzy subsets on the interval [-2MPa,2MPa] and [-20um,20um]. To apply nonsupervised fuzzy subset experiential estimation defines center point initialization and width of FRBF network. Thus according to 100 groups local site dada and η is 0.8, adopts above selection to obtain 45 center points.

Given that weights initialization values are 0.1, β is 0.7, sample time Ts is 0.001s and γ is 1, to apply above supervised learning algorithm to train weights.

To take into account entry strip gauge H is 3mm, entry strip crown degree Hd is 0.01mm, transverse tension deviation of entry strip σ_0 is 0.1Mpa and disturbance values are sine wave, namely,

$$\begin{cases} \Delta H = 0.02 \sin wt & (\text{mm}) \\ \Delta H_d = 0.01 \sin wt & (\text{mm}) \\ \Delta \sigma_0 = \sin wt & (\text{MPa}) \end{cases}$$

when this system expected output is as follows:

$$\begin{cases} \sigma_1 = 0.1 & (\text{MPa}) \\ h = 1.35 & (\text{mm}) \end{cases}$$

Simulation result is shown in Fig.3.a, Fig.3.b.

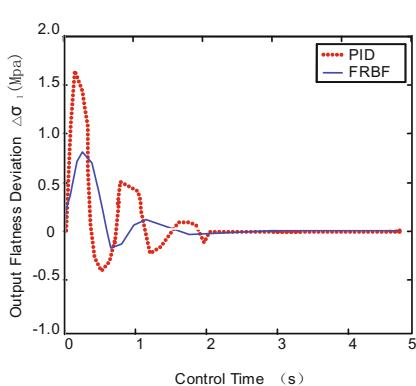


Fig. 3a. Automatic Gauge Control of Complex System

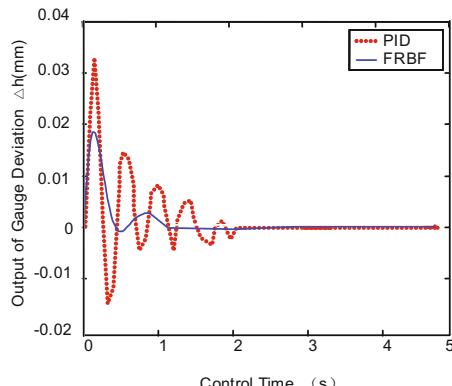


Fig. 3b. Automatic Flatness Control of Complex System

The simulations show that this FRBF multivariable control system can make flatness and gauge converge at set-up values without error, finely overcome the disturbance coming from entry strip, consequently eliminate the deviation of exit gauge. Meanwhile FRBF control system is compared to decoupled PID control, and results indicate that it has less oscillatory and extra- regulator value. Here in decoupled PID, flatness control parameters is respectively regulated to $K_p=-6e-6$, $T_i=3.5Ts$, $T_d=2.5Ts$, and gauge control parameters is $K_p=0.007$, $T_i=2.8Ts$, $T_d=1.5Ts$.

5 Conclusion

This flatness and gauge complex control system based on FRBF network performs the control task of first stand well, and effectively solves the problem that it is tough to control flatness and gauge with strongly coupled and nonlinear. Simulation results indicate that control effect is superior to conventional decoupled PID control. Consequently this control method establishes the foundation for further research about flatness and gauge in cold tandem rolling process.

References

- [1] Sun, Y.: Model and Control of Plate Strip Rolling at Cold - Hot Tandem Mill Rolling. Metallurgical Industry Press, Beijing (2010)
- [2] Wang, G.: Flatness Control and Flatness Theory. Metallurgy Industry Press, Beijing (1986)
- [3] Bao, H., Huang, X., Li, X., et al.: Control Theory and Application 17(2) (2000)
- [4] Kwasny, S.C., Faisal, K.A.: Rule-based training of neural networks. Expert Systems with Applications 2, 47–58 (1991)
- [5] Xu, L.: Neural Network Control. Harbin Institute of Technology Press, Harbin (1998)

Medical Image Segmentation Based on FCM and Wavelets

Zhihong Shi, Yi Liu^{*}, and Qian Li

School of Computer Science and Technology, Shandong University,
Jinan, 250101, China
liuyi@sdu.edu.cn

Abstract. In this paper, an unsupervised multiresolution image segmentation algorithm is put forward, which combines wavelet transform and improved fuzzy c-means clustering (FCM) considering neighboring pixels. In the first phase, the traditional FCM is applied to low-resolution image to get the initial image segmentation; Then, according to the properties of intrascale clustering and interscale persistence of wavelet coefficients, attach labels to image elements from coarse to fine scale; In the second phase, an improved FCM based on the neighboring pixels and obtained label of fine-scale, will be adopted for final image segmentation. In the experiments, medical images are segmented, which demonstrates the proposed method greatly restrains the influence of noise and shows good performance in the real medical images.

Keywords: FCM, Wavelet transform, Image segmentation, Multiresolution.

1 Introduction

Image segmentation is a very fundamental and vital step in image processing and computer vision. With the development of medical imaging, image segmentation plays an increasingly important role in medical image analysis [1,2]. It is a key step to extract quantitative information of the special organization in the medical image. But the complexity and diversity of medical images also bring difficulties to medical image segmentation. Compared with common image, medical images have heterogeneity, partial volume effect and noise. Therefore organizational boundaries label cannot be regarded as belonging to a single area, which must be determined by membership function. Fuzzy c-means clustering (FCM) [3,4] is a powerful tool to solve the above problem. It can assign one pixel to several different regions, with respect to fuzzy membership. FCM is an unsupervised algorithm and has been widely applied in image segmentation [5].

However, when FCM is applied to medical image segmentation, there are also several obvious drawbacks: 1) FCM algorithm is sensitive to initial value and is easy to fall into local minimum. So it is difficult to get the global optimal solution; 2) FCM

^{*} Corresponding author.

algorithm is over-sensitive to noise points, while the noise is inevitable in the medical image; 3) FCM algorithm does not take into consideration any spatial dependence. However, in real image pixels have strong local dependencies.

In order to overcome the shortcomings of the standard FCM algorithm, [6-8] added membership and scale parameter to the objective function of the FCM, proposed PCM algorithm. The algorithm greatly restrains the influence of noise, but how to determine the scale parameter of the algorithm is not easy. [9, 10] proposed a multi-resolution FCM algorithm. It not only reduces the sensitivity on the initialization, the whole property is also obviously increased than a single resolution. Moreover it is not sensitive to noise. However, complexity of the algorithm significantly increases. Inspired by these, this paper presents improved FCM algorithm combined with wavelet transform. Firstly, apply the FCM to low-resolution image to get the initial image segmentation; Secondly, we construct a novel fuzzy c-means clustering based on neighboring pixels and obtained label of fine-scale to complete the image segmentation.

2 Fuzzy C-Means Clustering (FCM)

FCM was first proposed by Dunn[3] and promoted by Bezdek [4].The core idea of FCM is to find the appropriate membership and cluster centroids, and make the fuzzy clustering objective function minimized, which can be defined as follows:

$$J(U, V) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^m d_{ik}(x_k, v_i) \quad (1)$$

In the above equation, $J(U, V)$ represents the sum of squares of distance between the pixels and the cluster centroid, the smaller its value, the better compactness of the image area, the better the clustering effect. c is the number of clusters, and N is the number of pixels. $X = \{x_1, x_2, \dots, x_n\}$ represents feature vector of each pixel; $v = \{v_1, v_2, \dots, v_c\}$ is the cluster centroid; $u = \{u_{ik}\}_{cn}$ is the membership matrix, u_{ik} is the membership of x_i is to k th cluster, which satisfies $\sum_{i=1}^c u_{ik} = 1$ and $u_{ik} \in [0, 1]$; m is weighted index, which determines the fuzzy degree of classification results. the greater its value, classification is more blurred, and it is generally 2; $d_{ik}(x_k, v_i)$ is distance of pixels to the cluster center, defined as follows:

$$d_{ik}(x_k, v_i) = \|x_k - v_i\|^2 \quad (2)$$

By Lagrange multiplier obtained the membership and cluster centroid iteratively update formula below:

$$u_{ik} = \left(\sum_{j=1}^c \left(\frac{d(x_k, v_i)}{d(x_k, v_j)} \right)^{2/(m-1)} \right)^{-1} \quad (3)$$

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m} \quad (4)$$

Update (3) (4), until $|J^{(l+1)} - J^l| \leq \epsilon$, to get the optimal solution. Then according to max membership, classify the pixels. If $u_{ji} > u_{jk}$, x_j is to the i th cluster, $k = 1, 2, \dots, c; i \neq k$.

3 The Proposed Method

When FCM is applied to the image segmentation, due to the single resolution and the effect of noise, causes over-segmentation. With the development of the theory of multi-scale and wavelet transform, multi-resolution FCM is more and more compelling. Its basic idea is to format the image cone of the different resolution by sampling and filtering of the image, apply FCM to different resolution, and finally fuse segmentation results in every resolution. This paper combines the wavelet transform and FCM to complete image segmentation. In the first part, after wavelet decomposition, the image generates image cone of different resolution. Then apply FCM to low-frequency image of the coarse-scale, label for each cluster and attach label to each pixel from a coarse scale to a fine scale. In the second part, we make use of the label in the first part and the neighboring pixels to improve standard FCM algorithm, and apply it to original image to get the final segmentation result. This method has the following advantages:

- 1) Because low-frequency image of the coarse scale have the global information of the image and a small amount of data, meanwhile is not sensitive to noise, the best initial segmentation results can be obtained in the shortest possible time.
- 2) In fine scale segmentation, our method makes full use of the coarse-scale information and neighboring pixels, so having more strong noise immunity and more accurate than single-resolution segmentation.

As a result, our proposed method will improve the efficiency greatly, as will be illustrated in the following subsections.

3.1 Wavelet Transform

Wavelet analysis is a new technique of time -scale analysis and multi-resolution analysis. Image after the wavelet transform is shown in Fig.1 (b). It is the wavelet coefficients distribution image after the classification of the three-scale. We try to get

an initial segmentation by the characteristics of the wavelet coefficients, which can be divided into two phases: firstly, apply the FCM to the coarse-scale. Secondly, reconstruct the image from the coarse-scale to the fine-scale.

Initial Segmentation. From characteristic of the wavelet analysis, we can learn that the low-frequency image of the coarse-scale has a small amount of data, is not sensitive to noise, but also has the global information and most of the energy of image. Therefore, we apply tradition FCM to low-frequency image of the coarse-scale to get the image initial segmentation result and the cluster centroids. Then according to the cluster result, we label for each cluster. So each pixel obtains only a label.

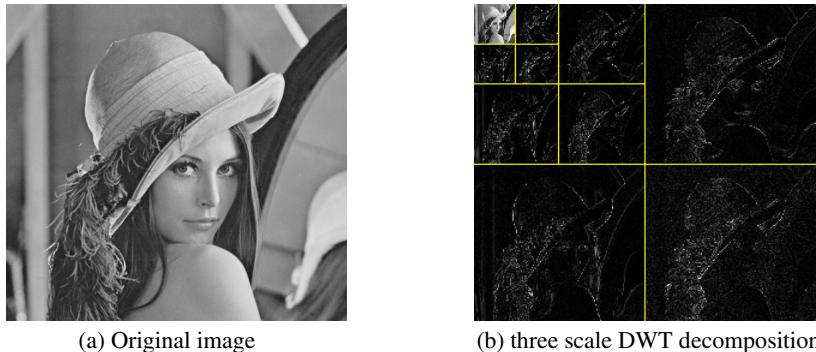


Fig. 1. Image DWT decomposition

Wavelet Reconstruction. According to the properties of intrascale clustering and interscale persistence of wavelet coefficients, we attach labels to image elements of each layer. For the quad-tree structure of wavelet decomposition [10], this is performed by evaluating the child-parent relationship in each layer. The child-parent relationship was validated by a similarity measure that takes the gray-value statistics of potential parent and a child into account. Among the potential parents, the label of a parent with the highest similarity is assigned to the child. The similarity between a child in layer n and a potential parent in layer $n-1$ is defined as

$$sim(x_n) = \frac{\sigma_{n-1} - (g(x_n) - \bar{x}_{n-1})^2}{k} \quad (5)$$

Where $g(x_n)$ is the gray value of child x_n in layer n . \bar{x}_{n-1} and σ_{n-1} are the average and standard deviation of the gray values of a region that a potential parent belongs to. The number of image elements of that region is denoted as k . A child inherits the label of the parent having maximum value of similarity. This completes labeling from the coarse scale to a fine scale, and finally gets the label of the most fine-scale images. Meanwhile we can get the initial cluster centroids of the image.

3.2 WFCM—An Improved FCM Based on Wavelets

Initial segmentation did not take advantage of the high-frequency coefficients of each layer, so the reconstructed image misses details. This paper using the information of reconstructed image to improve the FCM algorithm, and applied to the original image to obtain the result of having multi-resolution image segmentation.

Each pixel has a label from the reconstructed image, when two pixels' labels are same, the similarity of two pixels increases, and the distance of two pixels is reduced. We define a parameter, so that when the labels are same, the distance between the two pixels is reduced, or the distance is constant. The parameter is as follows:

$$\alpha_{i,j} = \begin{cases} t & x_i \text{ and } v_j \text{ are of the same labels} \\ 1 & x_i \text{ and } v_j \text{ are of the different labels} \end{cases} \quad (6)$$

In the above equation, $0 < t \leq 1$, in this paper, $t = 0.8$.

Defined the distance between two pixels:

$$d_{i,j}(x_i, v_j) = \|\alpha^{\lambda}_{i,j} \cdot (y_j - y_i)\|^2 \quad (7)$$

Where $\alpha_{i,j}$ represents the similarity between x_i and the cluster centroid v_j , λ is the degrees of freedom parameter, which determines the degree of influence that segmentation results of coarse resolution is to the distance d , this paper $\lambda = 1$.

y_i, y_j are the pixels' vector value.

Traditional FCM algorithm, only consider the pixel own gray value, no use of the information of the neighboring pixels. This paper redefine the pixel value, which is the weighted sum of the pixel own gray value and the average grays value of neighboring pixels. It is defined as follows:

$$y_i = \beta \cdot x_i + (1 - \beta) \cdot \frac{1}{N_R} \sum_{j \in N(x_i)} x_j \quad (8)$$

x_i is the pixel's gray value, where N_R is the number of neighboring pixels, and $N(x_i)$ is the neighboring pixels of the pixel x_i . $\frac{1}{N_R} \sum_{j \in N(x_i)} x_j$ represents the average grays value of neighboring pixels. $\beta (0 \leq \beta \leq 1)$ is the weighting factor, which is defined as follows:

$$\beta = \frac{1}{N_R} \sum_{j \in N(s_i)} s_{i,j} \quad (9)$$

N_R is the number of neighboring pixels, and $N(s_i)$ are the neighboring pixels of the pixel x_i . $s_{i,j}$ represents the similarity between x_i and x_j .

$$s_{i,j} = \begin{cases} 1 & |x_i - v_j| \leq T \\ 0 & |x_i - v_j| > T \end{cases} \quad (10)$$

Where T is a predefined threshold value. In our implementation the difference between the maximum and the minimum value of pixels was taken into account. If we denote this difference as diff , then the threshold was defined as

$$T = \text{diff} * \alpha; \quad \alpha \in (0,1) \quad (11)$$

Where α is the threshold factor.

Therefore, β represents the similarity between the pixel and its neighboring pixels. The more similar with the neighboring pixels, the larger β is.

The improved FCM algorithm in this paper can be described as follows:

Step 1. Apply the traditional FCM to the low-frequency image of the coarse scale. And obtain the cluster centroids and the initialized labels.

Step 2. Reconstruct the coarse image to the same size of original image. And get every pixel's label and the cluster centroids.

Step 3. Compute every pixel's feature value by (8).

Step 4. Update the distance by (7).

Step 5. Update the pixel membership based on (3).

Step 6. Update cluster centroids according to (4).

Step 7. Compute the objective function J_{new} by (1).

Step 8. If $|J_{\text{new}} - J_{\text{old}}| < \epsilon$, go to Step 9; else, go to Step 4.

Step 9. Perform image segmentation according to the membership of pixels.

4 Experimental Results and Analysis

This section will use the image in the first column of Fig.2 to illustrate the implementation of WFCM, and compare the experimental results of WFCM to traditional FCM. The first image is an abdomen image, whose size is 461×607 ; The second is a cancer of the liver image and the last is the second image with Gaussian white noise of mean 0 and variance 0.01. The sizes of them are 223×301 . We denote the three images simply by abdomen, liver, noisy image. The parameters in the experiments are set as follows: $c = 4$, $m = 2$, $\epsilon = 0.00001$ and the maximum number of iterations is assigned 100.

Fig.2 shows the segmentation results. Fig.2 (a) are the original images. Fig.2 (b) are the segmentation results of traditional FCM. Fig.2 (c) are the segmentation results of the improved FCM named NFCM, which only use neighboring pixels and no use wavelet transform. Fig.2 (d) are the segmentation results of WFCM. In the Fig.2, we

can see that FCM cannot overcome the degradation caused by noise in the segmentation performance, while WFCM completely succeeds in the images. The segmentation results of abdomen and liver show that in the segmentation images using FCM, there are many scattered dots and organization of misclassification is quite severe. The main reason for this is that FCM algorithm does not have space constraints capacity. However in the images of the NFCM and WFCM algorithm, the organization misclassification significantly reduced. The noisy images show that NFCM and WFCM performs better than traditional FCM for the image with noise, which demonstrates NFCM and WFCM can greatly restrain the influence of noise in the medical image segmentation. Compared with NFCM, WFCM is more accurate in the details of segmentation results. This is because that WFCM take full advantage of each level of information of wavelet decomposition. In summary, WFCM performs better than FCM and NFCM in the medical image segmentation.

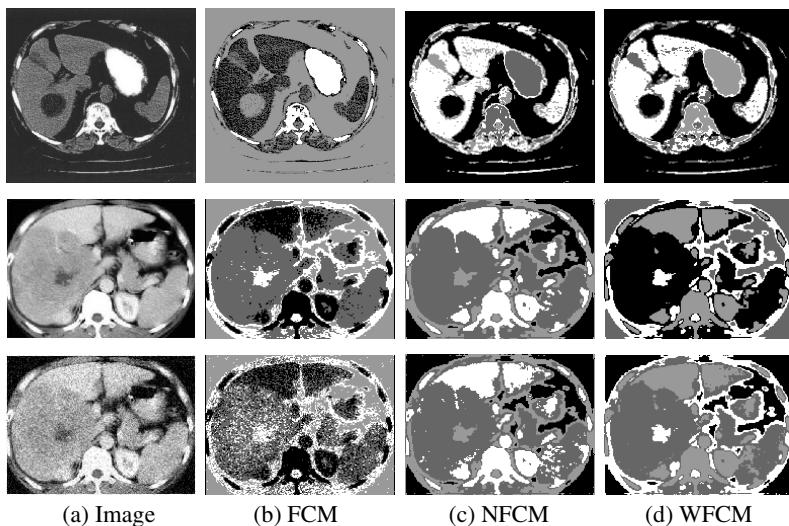


Fig.2. Segmentation Results

5 Conclusion

This paper presents an improved FCM algorithm based on wavelet transform, which takes advantages of the multiresolution properties of wavelet and advantage of FCM. Compared with the traditional FCM, our method greatly restrains the influence of noise and shows good performance in the real medical images. However, as segmentation method only using color information and neighboring pixels provided by the given image, the proposed method is insufficient for the segmentation of complex medical image. In future work we will solve the problem by finding more image future.

Acknowledgments. This work is supported by the Natural Science Foundation of Shandong Province (Grant Nos. ZR2011FM031).

References

1. Duncan, J.S., Ayache, N.: Medical Image Analysis: Progress over Two Decades and the Challenges Ahead. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22(1), 181–204 (2000)
2. Pham, D., Xu, C., Prince, J.: A Survey of Current Methods in Medical Image Segmentation. *Annual Review of Biomedical Engineering* 2(3), 315–337 (2000)
3. Dunn, J.C.: A graph theoretic analysis of pattern classification via Tamura's fuzzy relation. *IEEE Trans. Syst., Man Cybern.* 4, 310–313 (1974)
4. Bezdek, J.: *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York (1981)
5. Wang, X., Bu, J.: A fast and robust image segmentation using FCM with spatial information. *Digital Signal Processing* 20(4), 1173–1182 (2010)
6. Krishnaouram, R., Keller, J.M.: Possibilistic approach to clustering. *IEEE Transactions on Fuzzy System* 1(2), 98–110 (1993)
7. Krishnaouram, R., Keller, J.M.: The possibilistic c-means algorithm: Insights and recommendations. *IEEE Transactions on Fuzzy System* 4(3), 385–393 (1996)
8. Xie, Z.P., Wang, S.T., Zhang, D.Y., et al.: Image segmentation using the enhanced possibilistic clustering method. *Information Technology Journal* 6(4), 541–546 (2007)
9. Rezaee, M.R., Zwet, P.M.J.V.D., Lelieveldt, B.P.F., et al.: A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering. *IEEE Transactions on Image Processing* 9(7), 1238–1248 (2000)
10. Li, X.C., Bian, S.X.: Multiresolution fuzzy c-means clustering using markov random field for image segmentation. *International Journal of Information Technology and Computer Science* 1(1), 49–57 (2009)

A Derivative Augmented Lagrangian Method for Fast Total Variation Based Image Restoration

Dongwei Ren, Wangmeng Zuo, Hongzhi Zhang, and David Zhang

Biocomputing Research Centre, School of Computer Science and Technology

Harbin Institute of Technology, Harbin, 150001, China

{rendongwei@hit.edu.cn, cswmzuo, zhanghz0451}@gmail.com,
csdzhang@comp.polyu.edu.hk

Abstract. In this paper, we propose a novel derivative augmented Lagrangian method for fast total variation (TV) based image restoration (TVIR). By introducing a novel variable splitting method, TVIR is approximately reformulated in the derivative space, resulting in a constrained convex optimization problem which is simple to solve. Then, we propose a derivative alternating direction method of multipliers (D-ADMM) to solve the derivative space image restoration problem. Furthermore, we provide a Fourier domain updating algorithm which can save two fast Fourier transform (FFT) operations per iteration. Experimental results show that, compared with the state-of-the-art algorithms, D-ADMM is more efficient and can achieve satisfactory restoration quality.

Keywords: total variation, image restoration, augmented Lagrangian method, alternating direction method of multipliers, fast Fourier transform.

1 Introduction

Image restoration is known as a classic linear inverse problem [1], in which the latent image \mathbf{x} should be recovered from its degraded observation \mathbf{y} , modeled by

$$\mathbf{y} = \mathbf{Ax} + \mathbf{e}, \quad (1)$$

where \mathbf{A} is a linear degradation operator and \mathbf{e} is additive noise. Since the degradation operator \mathbf{A} usually is ill-conditioned, several regularizers, e.g., total variation (TV) [2], wavelet-based sparsity [3] and non-local model [4], have been proposed for image restoration. Because of its simplicity and robustness, TV regularizer has been widely applied into various image restoration applications, e.g., image denoising [5], blind deconvolution [6], and compressed sensing (CS) [7], and a number of methods have been proposed for TVIR.

On one hand, Augmented Lagrangian Method (ALM) is one class of the most efficient algorithms among various TVIR methods. Because of its non-smoothness, ALM-based methods for TVIR usually need incorporate some variable splitting strategies. In [8], by introducing a variable splitting strategy, Wang proposed several state-of-the-art fast TV deconvolution (FTVd) methods, including an alternating

minimization method in [8] and an alternating direction method of multipliers (ADMM) in [9]. In [10], Afonso et al. adopted another variable splitting strategy and developed a split augmented Lagrangian shrinkage algorithm (SALSA).

On the other hand, recently derivative space based formulation had also received considerable research interests in compressed sensing [11, 12], image restoration [13] and blind deconvolution [14]. In compressed sensing, Patel et al. showed that derivative space based approach can obtain higher success rate [11]. In image restoration, derivative space based method can work directly in the TV functional [13]. Besides, in image deconvolution, recent studies indicate that, the derivative space significantly outperforms the image space for the estimation of the blur kernel [3, 15].

In this paper, we unify these two directions by proposing a derivative space based ALM method for TVIR. First, we propose a novel approximate formulation of TVIR in the derivative space, providing an explanation of the derivative space based methods [11-13] from the viewpoint of variable splitting strategy. We then develop a D-ADMM algorithm to solve it, and provide a Fourier domain updating algorithm which can save two fast Fourier transform (FFT) operations per iteration. Compared with the state-of-the-art FTVd and SALSA algorithms, D-ADMM can achieve satisfactory restoration quality and is more efficient in terms of restoration speed.

The remainder of this paper is organized as: Section 2 introduces some preliminaries. Section 3 presents the proposed methods. Section 4 provides the experimental results. Finally, Section 5 ends this paper with some concluding remarks.

2 Prerequisites

In this section, we present some prerequisites used in latter context. Here a bold letter stands for a matrix or a vector, and if we arrange a matrix e.g., image \mathbf{x} , row by row into a vector, the same symbol \mathbf{x} will be used for saving notations.

2.1 TV-Based Image Restoration

Analogous to [13], we assume both the latent image \mathbf{x} and the degraded image \mathbf{y} lie in subspace \mathbb{U} with zero mean value, i.e., $\mathbb{U} = \{\mathbf{x} \in \mathbb{R}^{m \times n} \mid \text{mean}(\mathbf{x}) = 0\}$. The TVIR problem is formulated as,

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|^2 + \tau \text{TV}(\mathbf{x}), \quad (2)$$

where τ is a positive regularization parameter. There are usually two types TV regularizers, i.e., anisotropic and isotropic TV, defined as Eq. (3) and Eq. (4), respectively,

$$\text{TV}_a(\mathbf{x}) = \sum_{k=0}^{m-1} \sum_{l=0}^{n-1} \left(|(\mathcal{D}_h \mathbf{x})_{k,l}| + |(\mathcal{D}_v \mathbf{x})_{k,l}| \right), \quad (3)$$

$$\text{TV}_i(\mathbf{x}) = \sum_{k=0}^{m-1} \sum_{l=0}^{n-1} \sqrt{((\mathcal{D}_h \mathbf{x})_{k,l})^2 + ((\mathcal{D}_v \mathbf{x})_{k,l})^2}, \quad (4)$$

where the gradient operator $\mathcal{D} = \{\mathcal{D}_h, \mathcal{D}_v\}$, also notated as ∇ , is defined as

$$\begin{aligned} (\mathcal{D}_h \mathbf{x})_{k,l} &= \mathbf{x}_{k,l} - \mathbf{x}_{k,l-1}, \text{ with } \mathbf{x}_{k,-1} = \mathbf{x}_{k,n-1} \\ (\mathcal{D}_v \mathbf{x})_{k,l} &= \mathbf{x}_{k,l} - \mathbf{x}_{k-1,l}, \text{ with } \mathbf{x}_{-1,l} = \mathbf{x}_{m-1,l} \end{aligned} \quad (5)$$

where $k = 0, 1, 2, \dots, m-1$ and $l = 0, 1, 2, \dots, n-1$. Corresponding to these operators, there are matrices \mathbf{D}_h and \mathbf{D}_v such that $\mathbf{D}_h \mathbf{x} = \mathcal{D}_h \mathbf{x}$, and $\mathbf{D}_v \mathbf{x} = \mathcal{D}_v \mathbf{x}$. Accordingly, the adjoint operators \mathcal{D}_h^* and \mathcal{D}_v^* are associated with matrices \mathbf{D}_h^T and \mathbf{D}_v^T , respectively.

2.2 Moreau Proximal Mappings

TV-based image restoration usually involves the solution to some subproblems with the general form like,

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{w}\|^2 + \varepsilon g(\mathbf{x}), \quad (6)$$

where g is a (nonsmooth) convex function.

When g is l_1 norm, i.e., $g(\mathbf{x}) = \|\mathbf{x}\|_1$, the solution is

$$\mathcal{T}(w) = \operatorname{sgn}(w) \max(|w| - \varepsilon, 0) \quad (7)$$

where $T_\varepsilon(w) = \operatorname{sgn}(w) \max(|w| - \varepsilon, 0)$ is the soft-thresholding operator.

When g is $l_{2,1}$ norm, i.e., $g(\mathbf{x}) = \|\mathbf{x}\|_{2,1} = \sum_{l=0}^{N-1} \|\mathbf{x}_l\|_2$, then l -th column of solution is

$$\mathbf{x}_l = \mathcal{T}_\varepsilon\left(\|\mathbf{x}_l\|_2\right) \frac{\mathbf{w}_l}{\|\mathbf{w}_l\|_2}. \quad (8)$$

In our work, the derivative vector $\mathbf{x} = (\mathbf{x}_h^T, \mathbf{x}_v^T)^T$ is required to lie in the subspace \mathbb{V} of curl-free vector fields [13], and thus the image can be estimated from its gradient. Accordingly, the vector \mathbf{w} has two components \mathbf{w}_h and \mathbf{w}_v . Then we define a function

$$\iota_{\mathbb{V}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathbb{V} \\ +\infty, & \text{otherwise} \end{cases}. \quad (9)$$

When $g = \iota_{\mathbb{V}}$, the solution to (6) can be obtained by defining the projection $\nabla \mathcal{U}$ [13],

$$\mathbf{x} = \nabla \mathcal{U}(\mathbf{w}) = \nabla \text{FFT}^{-1} \left(\text{FFT}(\operatorname{div}(\mathbf{w})) \odot \mathbf{Wi} \right), \quad (10)$$

where \odot denotes entry-wise multiplication, the divergence is defined as

$$\operatorname{div}(\mathbf{w}) = -(\mathbf{D}_h^T \mathbf{w}_h + \mathbf{D}_v^T \mathbf{w}_v), \quad (11)$$

and the matrix \mathbf{Wi} with $\mathbf{Wi}(0,0) = 0$ and

$$(\mathbf{Wi})_{k,l} = 2\cos(2\pi k/m) + 2\cos(2\pi l/n) - 4, \quad (12)$$

where $k = 0, 1, 2, \dots, m-1$ and $l = 0, 1, 2, \dots, n-1$. Furthermore, we can also define the operator \mathcal{U} to estimate an image from its derivative vector,

$$\mathcal{U}(\mathbf{x}) = \text{FFT}^{-1} \left(\text{FFT}(\operatorname{div}(\mathbf{x})) \cdot \mathbf{Wi} \right). \quad (13)$$

3 The Derivative Augmented Lagrangian Method

In this section, we first present the formulation of the proposed derivative method, and then we solve it using ADMM.

3.1 Reformulation of TVIR

According to [16], if $\mathbf{u} \sim \mathcal{N}(\mu, \Sigma)$ where \mathcal{N} is Gaussian distribution, given the matrix \mathbf{A} , $\mathbf{Au} \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$. By assuming the unknown noise \mathbf{e} is with Guassian distribution, we have $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

Given the derivative operator $\mathbf{D} = \begin{pmatrix} \mathbf{D}_h \\ \mathbf{D}_v \end{pmatrix}$ with $\mathbf{De} = \begin{pmatrix} \mathbf{e}_h \\ \mathbf{e}_v \end{pmatrix}$, we can obtain $\begin{pmatrix} \mathbf{e}_h \\ \mathbf{e}_v \end{pmatrix} \sim \mathcal{N}(0, \sigma^2 \mathbf{DD}^T)$. Based on the definition of covariance matrix, we have,

$$E(\|\mathbf{e}\|_2^2) = mn\sigma^2 \quad \text{and} \quad E\left(\left\|\begin{pmatrix} \mathbf{e}_h \\ \mathbf{e}_v \end{pmatrix}\right\|^2\right) = \sigma^2 \text{tr}(\mathbf{DD}^T) = 4mn\sigma^2. \quad (14)$$

So we assume that $\|\mathbf{ADx} - \mathbf{Dy}\|^2 = 4\|\mathbf{Ax} - \mathbf{y}\|^2$ holds. Let $\mathbf{d} = \mathbf{Dx}$ with $\mathbf{d}_h = \mathbf{D}_h \mathbf{x}$ and $\mathbf{d}_v = \mathbf{D}_v \mathbf{x}$. The constraint $\mathbf{Dx} = \mathbf{d}$ requires that \mathbf{d} should lie in the subspace \mathbb{V} . Thus we approximately reformulate anisotropic TVIR as,

$$\mathbf{d} = \arg \min_{\mathbf{d} \in \mathbb{V}} \frac{1}{2} \|\mathbf{Ad} - \mathbf{a}\|^2 + \infty \|\mathbf{d}\|, \quad (15)$$

where $\infty = 4\tau$ and $\mathbf{a} = \mathbf{Dy}$ with $\mathbf{a}_h = \mathbf{D}_h \mathbf{y}$ and $\mathbf{a}_v = \mathbf{D}_v \mathbf{y}$. Numerical results show that, $\|\mathbf{Ad} - \mathbf{a}\|^2 / \|\mathbf{Ax} - \mathbf{y}\|^2 \sim \mathcal{N}(4.00, 1.62 \cdot 10^{-4})$, and thus it is reasonable to set $\infty = 4\tau$.

3.2 D-ADMM for Anisotropic TVIR

To solve the problem Eq. (15), we hereby propose a novel variable splitting strategy, introducing two auxiliary variables $\mathbf{f} = \mathbf{d}$ and $\mathbf{g} = \mathbf{d}$, and then Eq. (15) can be rewritten as,

$$\mathbf{d} = \arg \min_{\mathbf{d}, \mathbf{f}, \mathbf{g}} \frac{1}{2} \|\mathbf{Ad} - \mathbf{a}\|^2 + \infty \|\mathbf{f}\|_1 + \iota_{\mathbb{V}}(\mathbf{g}) \quad \text{s.t. } \mathbf{d} = \mathbf{f}, \mathbf{d} = \mathbf{g}, \quad (16)$$

which can be solved efficiently via ALM. The augmented Lagrangian function of Eq. (16) is first defined as,

$$\mathcal{L} = \infty \|\mathbf{f}\|_1 + \iota_{\mathbb{V}}(\mathbf{g}) + \frac{1}{2} \|\mathbf{Ad} - \mathbf{a}\|^2 + \frac{\delta_1}{2} \|\mathbf{d} - \mathbf{f} + \mathbf{p}\|^2 + \frac{\delta_2}{2} \|\mathbf{d} - \mathbf{g} + \mathbf{q}\|^2. \quad (17)$$

where the parameters \mathbf{p} and \mathbf{q} are associated to the Lagrangian multipliers. Then we can obtain the solutions to the subproblems with respect to \mathbf{d} , \mathbf{f} and \mathbf{g} within ADMM.

Given \mathbf{f} , \mathbf{g} , \mathbf{p} , and \mathbf{q} , with the help of FFT, the closed-form solution to \mathbf{d} can be obtained by

$$\mathbf{d} = FFT^{-1} \left(FFT \left(\mathbf{A}^T \mathbf{a} + \delta_1 (\mathbf{f} - \mathbf{p}) + \delta_2 (\mathbf{g} - \mathbf{q}) \right) \oslash FT_B \right), \quad (18)$$

where $FT_B = FFT(\mathbf{A}^T \mathbf{A}) + \delta_1 \mathbf{I} + \delta_2 \mathbf{I}$ and \oslash is entry-wise division. Given \mathbf{d} , \mathbf{g} , \mathbf{p} , and \mathbf{q} , the solution to \mathbf{f} can be obtained by,

$$\mathbf{f} = T_{\omega/\delta_1} (\mathbf{p} + \mathbf{d}). \quad (19)$$

Given \mathbf{d} , \mathbf{f} , \mathbf{p} , and \mathbf{q} , \mathbf{g} can be obtained by,

$$\mathbf{g} = \nabla \mathcal{U}(\mathbf{d} + \mathbf{q}). \quad (20)$$

Finally, the parameters \mathbf{p} and \mathbf{q} can be updated as follows

$$\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{d}_{k+1} - \mathbf{f}_{k+1} \quad \text{and} \quad \mathbf{q}_{k+1} = \mathbf{q}_k + \mathbf{d}_{k+1} - \mathbf{g}_{k+1}. \quad (21)$$

The penalty parameters δ_1 and δ_2 are fixed in conventional ADMM, leading to slow convergence rate. We hereby adopt the updating strategy in [19] to speed up convergence,

$$\delta_{1(k+1)} = \min(\delta_{\max}, \rho_1 \delta_{1(k)}) \quad \text{and} \quad \delta_{2(k+1)} = \min(\delta_{\max}, \rho_2 \delta_{2(k)}), \quad (22)$$

where δ_{\max} is upper bound of δ_1 and δ_2 , and ρ_1 and ρ_2 is defined as,

$$\rho_1 = \begin{cases} \rho_{1(0)}, & \text{if } \delta_1 \|\mathbf{d}_{k+1} - \mathbf{d}_k\| / \|\mathbf{f}_{k+1}\| < \varepsilon_1 \\ 1, & \text{otherwise} \end{cases} \quad \text{and} \quad \rho_2 = \begin{cases} \rho_{2(0)}, & \text{if } \delta_2 \|\mathbf{d}_{k+1} - \mathbf{d}_k\| / \|\mathbf{g}_{k+1}\| < \varepsilon_2 \\ 1, & \text{otherwise} \end{cases}, \quad (23)$$

where $\rho_{1(0)} > 1$ and $\rho_{2(0)} > 1$ are constants.

3.3 Fast D-ADMM in Fourier Domain

In D-ADMM, six FFT operations are required per iteration. Actually, \mathbf{d} , \mathbf{g} , \mathbf{p} , and \mathbf{q} can be updated in Fourier domain. By this way, only four FFT operations are required per iteration, resulting in a D-ADMM(F) algorithm.

First, let FT_b be the Fourier transform of $\mathbf{A}^T \mathbf{a}$, FT_f be the Fourier transform of \mathbf{f} , FT_g be the Fourier transform of \mathbf{g} , FT_p be the Fourier transform of \mathbf{p} , FT_q be the Fourier transform of \mathbf{q} . Then, the updating of \mathbf{d} can be performed in Fourier domain,

$$FT_d = (FT_b + \delta_1 (FT_f - FT_p) + \delta_2 (FT_g - FT_q)) \oslash FT_B \quad (24)$$

Then, we introduce the notation $\nabla \mathcal{U}_F$ in Fourier domain. Let FT_D_h and FT_D_v be the Fourier transform of \mathbf{D}_h and \mathbf{D}_v , respectively. Let $(FT_D_h)^*$ and $(FT_D_v)^*$ be the Fourier transform of the adjoint operators of \mathbf{D}_h and \mathbf{D}_v , respectively. The operator \mathcal{U}_F can be defined as

$$\mathcal{U}_F(FT_d) = -FFT^{-1}(((FT_D_h)^* \odot FT_d_h + (FT_D_v)^* \odot FT_d_v) \odot Wi), \quad (25)$$

and the projection $\nabla \mathcal{U}_F$ is defined as

$$\nabla \mathcal{U}_F(\text{FT_d}) = - \begin{pmatrix} \text{FT_D}_h \odot ((\text{FT_D}_h)^* \odot \text{FT_d}_h + (\text{FT_D}_v)^* \odot \text{FT_d}_v) \odot \text{Wi} \\ \text{FT_D}_v \odot ((\text{FT_D}_h)^* \odot \text{FT_d}_h + (\text{FT_D}_v)^* \odot \text{FT_d}_v) \odot \text{Wi} \end{pmatrix}. \quad (26)$$

By using $\nabla \mathcal{U}_F$, FT_g can be updated by

$$\text{FT_g} = \nabla \mathcal{U}_F(\text{FT_d} + \text{FT_q}). \quad (27)$$

With FT_d and FT_p , FT_f can be updated by

$$\text{FT_f} = FFT \left(\mathcal{T}_{\omega/\delta_1} (FFT^{-1}(\text{FT_p} + \text{FT_d})) \right). \quad (28)$$

Finally, we summarize D-ADMM in Fourier domain, i.e., D-ADMM(F), in Algorithm 1.

Algorithm 1: D-ADMM(F)

1. Preprocess $\bar{\mathbf{y}} = \text{mean}(\mathbf{y})$
 2. Initialize $\text{FT_p}_0, \text{FT_q}_0, \text{FT_d}_0, \text{FT_f}_0, \text{FT_g}_0, k = 0$
 3. Precompute $\text{FT_b} = FFT(\mathbf{A}^T \mathbf{Dy}),$

$$\text{FT_B} = FFT(\mathbf{A}^T \mathbf{A}) + \delta_1 \mathbf{I} + \delta_2 \mathbf{I}$$
 4. **while** not converged
 5. $\text{FT_d}_{k+1} = (\text{FT_b} + \delta_1(\text{FT_f}_{k+1} - \text{FT_p}_{k+1}) + \delta_2(\text{FT_g}_{k+1} - \text{FT_q}_{k+1})) \oslash \text{FT_B}$
 6. $\text{FT_f}_{k+1} = FFT(\mathcal{T}_{\omega/\delta_{1(k)}} (FFT^{-1}(\text{FT_p}_k + \text{FT_d}_{k+1})))$
 7. $\text{FT_g}_{k+1} = \nabla \mathcal{U}_F(\text{FT_d}_{k+1} + \text{FT_q}_k)$
 8. $\text{FT_p}_{k+1} = \text{FT_p}_k + \text{FT_d}_{k+1} - \text{FT_f}_{k+1}$
 9. $\text{FT_q}_{k+1} = \text{FT_q}_k + \text{FT_d}_{k+1} - \text{FT_g}_{k+1}$
 10. Update $\delta_{1(k+1)}$ and $\delta_{2(k+1)}$ using Eq. (22)
 11. $k = k + 1$
 12. **end while**
 13. $\mathbf{x} = \mathcal{U}_F(\text{FT_d}_k)$
 14. $\mathbf{x} = \mathbf{x} + \bar{\mathbf{y}}$
-

Furthermore, D-ADMM and D-ADMM(F) can be easily extended to isotropic TV by modifying the shrinkage operator [17].

3.4 Implementation Issues

Rather than stopping the program in a fixed number of iterations, we adopt the stopping criteria by checking the difference in the variable \mathbf{d}_k and \mathbf{d}_{k+1} is whether below a sufficient small positive value ε ,

$$\|\mathbf{d}_{k+1} - \mathbf{d}_k\| / \|\mathbf{d}_k\| \leq \varepsilon. \quad (29)$$

Both \mathbf{p}_0 and \mathbf{q}_0 are initialized to be zero. For fast convergence, we empirically give the following recommendation on the initialization of $\delta_{1(0)}, \delta_{2(0)}, \delta_{\max}, \rho_{1(0)}, \rho_{2(0)}, \varepsilon_1$, and ε_2 : $\delta_{1(0)} = \delta_{2(0)} = 10^{-4}$, $\delta_{\max} = 100 \max(\delta_{1(0)}, \delta_{2(0)})$, $\rho_{1(0)} = 2.5$, $\rho_{2(0)} = 1.9$, and $\varepsilon_1 = \varepsilon_2 = 10^{-3}$.

4 Experimental Results

In this section, we use five 256×256 images, i.e., Lena, Cameraman, Barbara, Baboon and Couple, to evaluate the efficiency and effectiveness of the proposed algorithm for isotropic TVIR. We compare D-ADMM with two state-of-the-art ALM-based TVIR methods, i.e., SALSA [10] and FTVd [9].

In the experiments, each image is blurred by 9×9 Gaussian kernel with the standard deviation (*std.*) of 4, and noised by normally distributed noise with mean of zero and *std.* of 10^{-3} . As to parameter setting, we choose the value $\varepsilon = 10^4$, the regularization parameters as $\mu = 5 \times 10^{-5}$ and $\tau = \mu / 4$. For performance evaluation, we adopt peak signal-to-noise ratio (PSNR) and complex wavelet structural similarity (SSIM) [18] to assess the restoration quality, and the CPU run time to evaluate the restoration speed.

Table 1. Results of comparative experiments: t stands for CPU time (s), p stands for PSNR, and s stands for SSIM

method	Lena t/p/s	Cameraman t/p/s	Barbara t/p/s	Baboon t/p/s	Couple t/p/s
SALSA [10]	9.73/31.81/0.90	9.92/31.27/0.92	7.86/31.00/0.87	5.27/26.34/0.77	8.14/32.07/0.91
FTVd [9]	0.90/31.81/0.90	1.09/31.30/0.92	1.01/30.96/0.87	0.75/26.35/0.77	1.01/31.95/0.91
D-ADMM	0.83/31.35/0.90	1.03/30.73/0.90	1.05/30.96/0.86	0.91/26.12/0.75	1.13/31.65/0.91
D-ADMM(F)	0.70/31.46/0.90	0.59/30.85/0.91	0.55/31.06/0.87	0.51/26.33/0.76	0.59/31.75/0.91

To save space, we only shows the restoration result of Barbara using D-ADMM(F) in Figure 1. Table 1 lists the run time (t), PSNR (p), and SSIM (s) obtained using SALSA [10], FTVd [9], D-ADMM, and D-ADMM(F). From Table 1, D-ADMM and D-ADMM(F) are comparable with SALSA and FTVd in terms of both PSNR and SSIM. In terms of CPU run time, D-ADMM(F) is more efficient than SALSA and FTVd. Meanwhile, D-ADMM(F) with lower complexity per iteration is faster than D-ADMM and FTVd.

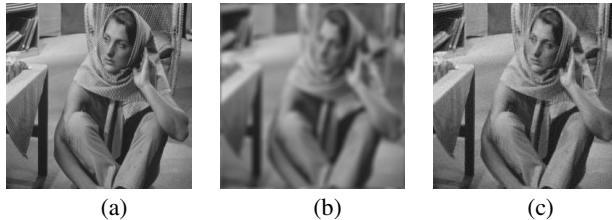


Fig. 1. Restoration result of Barbara by D-ADMM(F). a) original image, b) degraded image, and c) restoration result.

5 Conclusion

In this paper, we present a novel viewpoint on the reformulation of TVIR in derivative space and a novel ALM-based algorithm, i.e., D-ADMM. Based on probabilistic analysis, we approximately reformulate TVIR into the derivative space, resulting in a simpler constrained convex optimization problem. To solve this problem, we develop a novel

D-ADMM algorithm, and further propose a D-ADMM(F) algorithm to directly update in Fourier domain. Finally, experimental results indicate that, compared with SALSA and FTVd, D-ADMM(F) is more efficient and can achieve comparable restoration quality.

Acknowledgement. The work is partially supported by the NSFC funds of China (Grant No.s: 61271093, 61001037, and 61071179).

References

- [1] Andrews, H., Hunt, B.: *Digital Image Restoration*. Prentice-Hall, Englewood Cliffs (1977)
- [2] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60(1-4), 259–268 (1992)
- [3] Donoho, D., Johnstone, I.: Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3), 425–455 (1994)
- [4] Peyré, G., Bougleux, S., Cohen, L.: Non-local Regularization of Inverse Problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III. LNCS*, vol. 5304, pp. 57–68. Springer, Heidelberg (2008)
- [5] Chambolle, A.: An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision* 20(1-2), 89–97 (2004)
- [6] Chan, T.F., Wong, C.K.: Total variation blind deconvolution. *IEEE Trans. IP* 7(3), 370–375 (1998)
- [7] Ma, S., Yin, W., Zhang, Y., Chakraborty, A.: An efficient algorithm for compressed MR imaging using total variation and wavelets. In: *CVPR* (2008)
- [8] Wang, Y., Yang, J., Yin, W., Zhang, Y.: A New Alternating Minimization Algorithm for Total Variation Image Reconstruction. *SIAM Journal on Imaging Sciences* 1(3), 248–272 (2008)
- [9] Tao, M., Yang, J., He, B.: Alternating direction algorithms for total variation deconvolution in image reconstruction. In: TR0918, Department of Mathematics, Nanjing University (2009)
- [10] Afonso, M.V., Bioucas-Dias, J.M., Figueiredo, M.A.T.: Fast Image Recovery Using Variable Splitting and Constrained Optimization. *IEEE Trans. IP* 19(9), 2345–2356 (2010)
- [11] Patel, V.M., Maleh, R., Gilbert, A.C., Chellappa, R.: Gradient-Based Image Recovery Methods From Incomplete Fourier Measurements. *IEEE Trans. IP* 21(1), 94–105 (2012)
- [12] Rostami, M., Michailovich, O., Zhou, W.: Image Deblurring Using Derivative Compressed Sensing for Optical Imaging Application. *IEEE Trans. IP* 21(7), 3139–3149 (2012)
- [13] Michailovich, O.V.: An Iterative Shrinkage Approach to Total-Variation Image Restoration. *IEEE Trans. IP* 20(5), 1281–1299 (2011)
- [14] Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: *ACM SIGGRAPH 2006*, pp. 787–794 (2006)
- [15] Dong, W., Zhang, L., Shi, G., Wu, X.: Image Deblurring and Super-Resolution by Adaptive Sparse Domain Selection and Adaptive Regularization. *IEEE Trans. IP* 20(7), 1838–1857 (2011)
- [16] Bishop, C.M.: *Pattern recognition and machine learning*. Springer, New York (2006)
- [17] Zuo, W., Lin, Z.: A Generalized Accelerated Proximal Gradient Approach for Total-Variation-Based Image Restoration. *IEEE Trans. IP* 20(10), 2748–2759 (2011)
- [18] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. IP* 13(4), 600–612 (2004)
- [19] Lin, Z., Liu, R., Su, Z.: Linearized alternating directional method with adaptive penalty for low-rank representation. In: *NIPS* (2011)

A Novel Multi-class Brain-Computer Interface (BCI) Paradigm Based on Motor Imagery Sequential Coding (MISC) Protocol

Jun Jiang, Erwei Yin, Yang Yu, Jingsheng Tang, Zongtan Zhou, and Dewen Hu

College of Mechatronics and Automation, National University of Defense Technology,
Changsha, Hunan 410073, P.R. China
dwhu@nudt.edu.cn

Abstract. In this study, we present a novel multi-class BCI paradigm based on motor imagery sequential coding (MISC) protocol, which can generate multiple commands just by two kinds of motor imagery (MI) tasks. In the MISC protocol, each mental task was divided into several continuous epochs with the same duration. During each epoch, one of the two MI tasks was executed. With this protocol, multiple mental states can be coded by the two MI tasks. Additionally, the difficulty of classifier design was also reduced as only two MI tasks were needed to be classified. Three subjects participated in our experiments, and achieved an average accuracy of 85.7%, with the ITR of 16.5 bits/min. The results confirmed that the MISC protocol can generate more commands in BCI system with the equal number of MI tasks.

Keywords: Electroencephalogram (EEG), Sensorimotor Rhythms, Multi-class brain-computer interface, Motor imagery Sequential Coding.

1 Introduction

Brain-computer interface (BCI) is an alternative communication and control channel between humans and artificial devices, which can translate brain activities to commands without the participation of peripheral nerves and muscles. BCI can serve as an effective assistive technology for disabled individuals to restore their lost communication and motor function [1]. Recent years, both invasive and noninvasive BCI have made significant progress. In this study, we focus on the noninvasive BCI based on electroencephalograms (EEG) signals.

Sensorimotor rhythms (SMRs) are regarded as good features for EEG-based BCI, because it can be modulated by imagining kinesthetic movement without actual physical activities and show consistent patterns on the sensorimotor cortical areas [2, 3]. Furthermore, comparing with the other features of EEG, i.e. P300 and visual evoked potential (VEP) [4, 5], the manner to modulate SMRs is more active and voluntary which do not need external stimulations. The above advantages make the SMR-based BCI become a suitable assistant technique for the physical disabilities. To date, many kinds of SMR-based BCI

systems have been established and some of them have been applied to some practical applications, like computer cursor control, brain-actuated wheelchair and neuroprosthesis [6–8].

However, the lack of available mental states generated by MI tasks precludes the development of SMR-based BCI. In theory, multiple mental states can be realized by imagining different kinesthetic movements, such as left/right hand, foot, and tongue movement [9–11], due to these brain activities have different spatial localization of SMR modulations in sensorimotor cortex, which can inducing discriminative EEG patterns. However, in practice, to detect these mental states effectively, the signal processing and classifiers design become a critical challenge, as the low signal to noise ratio (SNR) of EEG. Although a number of algorithms have been proposed to try to solve this problem [10–13], the problem of overfitting and computational cost was also raised along with the increase of algorithm complexity. In fact, every additional EEG pattern to be classified will bring up more difficulty to the classifier design and thus there is a trade-off between the accuracy and the amount of MI tasks. Obermaier *et al.* (2001) evaluated the performance of a five-class BCI system. Their results showed that, to achieve the highest information transfer rate (ITR), the upper limit of different mental states for a BCI system is three [14].

In this paper, we propose a novel multi-class BCI paradigm based on motor imagery sequential coding (MISC) protocol, which can generate multiple output commands just by two kinds of basic MI tasks. In the MISC protocol, the basic MI tasks were sequentially executed in a fixed duration, and thus multiple mental states can be coded. To identify the two basic MI tasks, there are many of reliable methods [15, 16], which can reduce the difficulty of signal processing and classification algorithm in our paradigm. The proposed paradigm was tested in the offline experiments and the performance achieved an average classification accuracy of 85.7% with an ITR of 16.5 bits/min, which confirmed the effectiveness of our proposed multi-class BCI paradigm.

2 Materials and Methods

2.1 The MISC Protocol

Under the MISC protocol, the output command was produced by the sequential MI task, which was divided into several continuous epochs with the same duration T . During each epoch, subjects were instructed to execute one of two basic MI tasks. With this protocol, the sequential MI task was coded in a sense of permutation. According to the permutation theory, if N epochs are contained in one sequential MI task (called N -length sequential MI task), there are 2^N different kinds of mental states could be generated. Therefore, multiple commands could be produced by limited kinds of MI tasks, as well as the difficulty of classifier design was reduced as only two MI tasks were needed to be classified in the MISC protocol.

In our study, imagining left and right hand movement were selected as the basic MI tasks, because these two tasks are represented at the different hemisphere

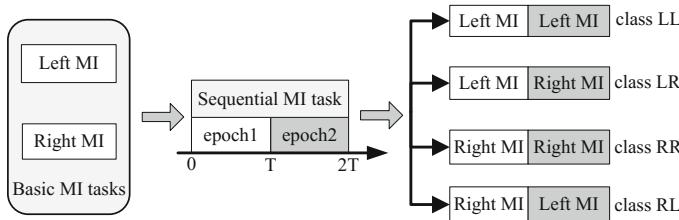


Fig. 1. The MISC protocol based on left/right hand MI tasks. A 2-length sequential MI task was designed in this paper, which consisted of 2 continuous epochs.

of the brain that the corresponding EEG patterns have a high discrimination. To show the feasibility of the MISC protocol, we applied the 2-length sequential MI tasks to realize a four-class BCI paradigm, as illustrated in Fig. 1.

2.2 Subjects and Data Acquisition

Three healthy right-handed subjects participated in the study. During the experiments, the subjects were seated in a comfortable armchair facing a 22 in. LCD computer screen. All subjects were informed about the detailed procedures of the experiments, and signed an informed consent form in accordance with the Declaration of Helsinki.

Sixteen-channel EEG data were recorded around the sensorimotor cortex (F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P3 and P4, referenced to P8 and grounded to Fpz) based on the international 10-20 system. The impedances of all the electrodes were kept below 10 k. Data were acquired by a BrainAmp DC Amplifier (Brain Products GmbH, Germany) with a sampling rate of 250 Hz.

2.3 Experimental Procedures

The experiments in our study consisted of two procedures. The first was the training experiment for left/right MI tasks. Both the classifiers for left/right MI tasks and the subjects' skill to modulate SMR efficiently were trained in this procedure. The second procedure was the recognition experiment for sequential MI tasks. The classification accuracy and ITR were evaluated from the collected data.

Training Experiment for Basic MI Tasks. A typical cue-based training paradigm which consisted of non-feedback and feedback training was performed in this experiment. In the non-feedback training, subjects were instructed to execute left/right MI task according to the literal cues (START, LEFT, RIGHT and PAUSE). A whole trial of one task last for 10 seconds. For the first 3 seconds, "START" was appeared on the screen to notify the beginning of the trial. Then,

one of the MI cues (LEFT or RIGHT) was maintained on the screen for 5 seconds, and the subjects tried to execute the corresponding task. To prevent forecasting, the literal cues appeared randomly. After that, a 2-second interval with "PAUSE" cue was allocated before the next trial (see Fig. 2(a)).

Feedback methods for motor imagery training can help subjects master the skill of modulate SMR efficiently and produce consistent EEG patterns. Therefore, a feedback training paradigm was introduced in the experiment. The framework of the feedback training was almost the same with non-feedback training, except that the literal cues could be moved to provide feedback information. The EEG signals were classified as left/right commands by the initial classifier to move the literal cues. Subjects were asked to perform the same quantity of mental tasks with non-feedback training, and the classification accuracy was evaluated. If the accuracy was at least 85%, the experiment was finished. Otherwise, the whole training experiment was repeated sequentially until the accuracy criterion was satisfied.

Recognition Experiment of Sequential MI Tasks. In this experiment, the four sequential MI tasks, labeled as "LL", "LR", "RR" and "RL" (see Fig. 1) were recognized from the ongoing EEG measurements. A whole trial of one task lasted for 15 seconds. For the first 6 seconds, "START" was appeared on the screen to notify the beginning of the trial. Then, one of four SMI cues ("LL", "LR", "RR" or "RL") was maintained on the screen for 6 seconds, and the subjects tried to execute the corresponding task. To prevent forecasting, the literal cues also appeared in random. During the MI period, to forbid the confusion of the switching time from epoch1 to epoch2, a time slider was displayed in the screen, to help subjects perform SMI tasks correctly (see Fig. 2(b)). After that, a 3-second interval with "PAUSE" cue was allocated before the next trial started. In this experiment, every subject was asked to perform 56 trials for each sequential MI task, and the collected data were analyzed offline to evaluate the performance of our multi-class BCI paradigm.

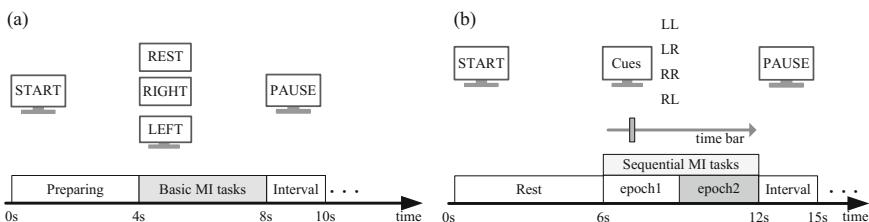


Fig. 2. (a) The non-feedback and feedback training experiment of the basic MI tasks; (b) The recognition experiment of sequential MI tasks. The time bar started to move at 6s and stopped at 12s, which can point out the switch time from epoch1 to epoch2.

2.4 Signal Processing

The process of signal processing in our multi-class BCI paradigm was organized as follows: first, EEG signals were classified as left/right commands; next, the sequence of left/right commands was compared to the four sequential MI tasks, and the four similarity measurements were calculated respectively; finally, the sequential MI with the optimal similarity measurement was exported. The details of the signal processing are described in the following subsections.

Classification of Basic MI Tasks. The common spatial pattern (CSP) and linear discriminant analysis (LDA) algorithms were used in this study to train the classifier of left/right MI tasks, as the two algorithms are widely used in SMR-based BCI system [15, 16].

All EEG channels were filtered between 8-30Hz and the filtered signals in MI period were divided into 1000ms slices. The spatial filter was trained from the data sets by CSP algorithm. An optimized sub-filter is selected from the CSP filter [15], and the logarithmic variance values were calculated on the filtered EEG signals to obtain a 2-dimension feature vectors. Finally, LDA was utilized to construct the classifier from the feature vectors.

Classification of Sequential MI Tasks. A template matching method was designed in this procedure, to recognize the four sequential MI tasks. In our method, the flag for left MI was -1 and for right MI was +1. According to the sequential MI tasks, four standard templates were designed as the criterion of each task (see Fig. 3).

In the experiment, the command sequences with the same length of the standard templates were extracted to recognize the four sequential MI tasks. For a given sequence sample $x = (x_1, x_2, \dots, x_n)$, the Euclidean distance $D_k(x)$ was calculated to evaluate the similarity measurement between and each standard template:

$$D_k(x) = \|x - M_k\|_2 = [\sum_{i=0}^n (x_i - m_k(i))^2]^{1/2} \quad (1)$$

where $k \in \{LL, LR, RR, RL\}$, and $M_k = (m_k(1), m_k(2), \dots, m_k(n))^T$ denotes the standard template. When the four Euclidean distances were obtained, the normalized distances $ND_k(x)$ were calculated by

$$ND_k(x) = D_k(x) / \sum_k (D_k(x)) \quad (2)$$

The range of $ND_k(x)$ is $[0, 1]$ and the lower value of $ND_k(x)$ indicates the higher similarity to the corresponding standard template. The final output of x is the class with the minimal $ND_k(x)$ provided that the distance is below a given threshold; otherwise the result is "unknown". The choice of appropriate threshold was guided by a receiver operating characteristic (ROC) analysis [17].

To evaluate the performance of our BCI paradigm, the accuracy and the ITR were both calculated [2]. The accuracy is defined as the ratio of the number of correctly matched trials to the total number of trials. The ITR was given by

$$ITR = \left\{ \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{N - 1} \right\} / T \quad (3)$$

Where P denotes the accuracy and T is the response time which defined as the time taken until the correct command is detected.

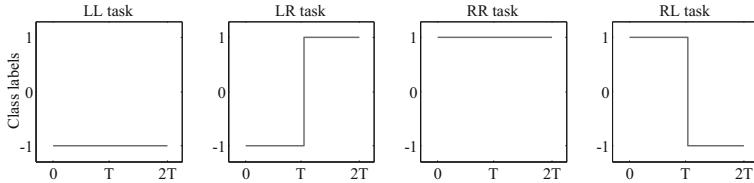


Fig. 3. The four standard templates of the sequential MI tasks

3 Results

In the recognition experiment, the time window length of the standard templates was 4s. Table 1 shows the performance of the experiment. The average accuracy and ITR were 85.7% and 16.5 bits/min respectively, which subject A got the highest values of 89.7% and 19.1 bits/min. The response time (RT) for each subject was also showed in Table 1. The mean RT was 4.34s which was slightly longer than the ideal RT (4s, with the same length of standard templates), because subjects need some reaction time to execute the tasks when the literal cues displayed at the beginning.

Table 1. Performance of the four sequential MI tasks recognition

Subject	Response time (s)	Threshold	Accuracy (%)					ITR (bits/min)
			LL	LR	RR	RL	Average	
A	4.27	0.12	92.9	89.3	94.6	82.1	89.7	19.1
B	4.32	0.15	91.1	78.6	91.1	80.4	85.3	16.2
C	4.42	0.13	89.3	75.0	87.5	76.8	82.2	14.1
Mean	4.34	0.13	91.1	81.0	91.1	79.8	85.7	16.5

The samples of the four sequential MI tasks were extracted from real EEG signals. All samples were averaged over 56 trials for each task, and compared to the standard templates. The four cursors of the samples had the similar form with the corresponding standard templates, as illustrated in Fig. 4. The cursors of "LL" and "RR" tasks matched the standard templates better than "LR" and "RL" tasks, which the mean normalized distance (see in section2.3.2) for "LL" and "RR" tasks was 0.02 and for "LR" and "RL" was 0.11.

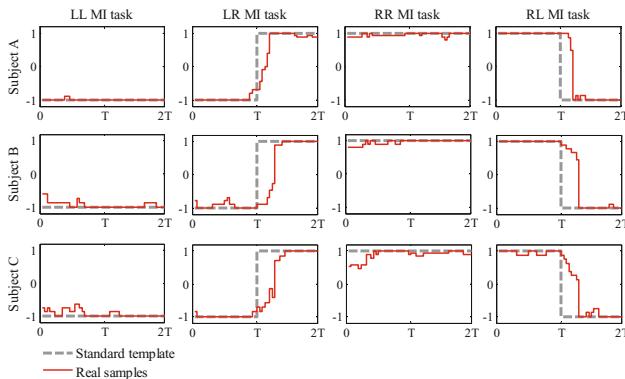


Fig. 4. The samples of the four sequential MI tasks extracted from real EEG signals

4 Conclusions

In this study, we proposed a novel multi-class BCI paradigm based MISC protocol. In the MISC protocol, a 2-length sequential MI tasks were designed with left and right MI tasks, to realize the four classes of brain states. The EEG signals were first classified as left or right commands using CSP and LDA algorithms. Then, the sequence of left/right commands was processed by the template matching method, to recognize the four sequential MI tasks. For all the subjects, the average accuracy of the four sequential MI tasks was 85.7% and the average ITR was 16.5 bits/min. The results confirmed that the MISC protocol can generate more commands in BCI system with the equal number of MI tasks. More importantly, as there were only two kinds of EEG patterns needed to be classified, the difficulty of signal processing and classifier design was reduced significantly.

In future work, we will perfect the MISC protocol to enhance the performance of BCI paradigm. At the same time, we will apply our BCI paradigm to controlling some external devices to make it to be a practical assistant technique for physical disabilities.

Acknowledgments. This work was supported in part by the National High Technology Research and Development Program (Project 2012AA011601) and the National Basic Program of China (Project 2011CB707802).

References

1. Dornhege, G., Milln, J.R., Hinterberger, T., McFarland, D.J., Mller, K.R.: *Toward Brain Computer Interfacing*. The MIT Press, London (2007)
2. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-Computer Interfaces for Communication and Control. *Clin. Neurophysiol.* 113, 767–791 (2002)

3. Pfurtscheller, G., Neuper, C.: Motor Imagery Activates Primary Sensorimotor Area in Humans. *Neuroscience Letters* 239, 65–68 (1997)
4. Piccione, F., Giorgi, F., Tonin, P.: P300-based Brain Computer Interface: Reliability and Performance in Healthy and Paralysed Participants. *Clin. Neurophysiol.* 117, 531–537 (2006)
5. Lee, P.J., Hsieh, J.C., Wu, C.H.: Brain Computer Interface Using Flash Onset and Offset Visual Evoked Potentials. *Clin. Neurophysiol.* 119, 605–616 (2008)
6. Wolpaw, J.R., McFarland, D.J.: Control of a Two-dimensional Movement Signal by a Noninvasive Brain-Computer Interface in Humans. *Proceedings of the National Academy of Sciences of United States of America (PNAS)* 101, 17849–17854 (2004)
7. Galn, F., Nuttin, M., Lew, E., Ferrez, P.W., Vanacker, G., Philips, J., Milln, J.R.: A Brain-Actuated Wheelchair: Asynchronous and Non-invasive Brain-Computer Interfaces for Continuous Control of Robots. *Clin. Neurophysiol.* 119, 2159–2169 (2008)
8. Mller-Putz, G.R., Scherer, R., Pfurtscheller, G., Rupp, R.: EEG-based Neuroprostheses Control: Step towards Clinical Practices. *Neuroscience Letters* 382, 169–174 (2005)
9. Lemm, S., Blankertz, B., Curio, G., Mller, K.R.: Spatio-spectral Filters for Improving the Classification of Single Trial EEG. *IEEE Trans. Biomed. Eng.* 52, 1541–1548 (2005)
10. Gouy-Pailler, C., Congedo, M., Brunner, C., Jutten, C., Pfurtscheller, G.: Non-stationary Brain Source Separation for Multi-class Motor Imagery. *IEEE Trans. Biomed. Eng.* 57, 469–478 (2010)
11. Dornhege, G., Blankertz, B., Curio, G., Mller, K.R.: Boosting Bit Rates in Non-invasive EEG Single-Trial Classifications by Feature Combination and Multiclass Paradigms. *IEEE Trans. Biomed. Eng.* 51, 993–1002 (2004)
12. Grosses-Wentrup, M., Buss, M.: Multiclass Common Spatial Patterns and Information Theoretic Feature Extraction. *IEEE Trans. Biomed. Eng.* 55, 1991–2000 (2008)
13. Brunner, C., Naeem, M., Leeb, R., Graimann, B., Pfurtscheller, G.: Spatial Filtering and Selection of Optimized Components in Four Class Motor Imagery EEG Data Using Independent Components Analysis. *Pattern Recognition Letters* 28, 957–964 (2007)
14. Obermaier, B.C., Neuper, C., Pfurtscheller, G.: Information Transfer Rate in Five-Classes Brain Computer Interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* 9, 283–288 (2001)
15. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Mller, K.R.: Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine* 25, 41–56 (2008)
16. Lotte, F., Congedo, M., Lcuyer, A., Lamarche, F., Arnaldi, B.: A Review of Classification Algorithms for EEG-Based Brain-Computer Interfaces. *J. Neural Eng.* 4, R1–R3 (2007)
17. Townsend, G., Graimann, B., Pfurtscheller, G.: Continuous EEG Classification During Motor Imagery-Simulation of an Asynchronous BCI. *IEEE Trans. Neural Syst. Rehabil. Eng.* 12, 258–255 (2004)

Image Segmentation Based on NSCT and BF-PSO Algorithm

Le Wang and Zhenbing Zhao

School of Electrical and Electronic Engineering, North China Electric Power University
Baoding, Hebei, China
heihuoyi@126.com, zhaozhenbing2002@163.com

Abstract. Image segmentation is an important part of image processing. To improve the quality and the speed of the segmentation, a new method based on Non Subsampled Contourlet Transform (NSCT) and Bacterial Foraging-Particle Swarm Optimization (BF-PSO) algorithm is proposed in the paper. In this method, a gray-gradient co-occurrence matrix based on NSCT is constructed. On the basis of the matrix, a gray entropy model is constructed. The fitness function based on the gray entropy model is designed for BF-PSO algorithm. Then, obtain the best threshold when the entropy reaches its maximum value, and then complete the segmentation. A group of experiments indicate that the proposed method is efficient and needs less iteration.

Keywords: segmentation, NSCT, co-occurrence matrix, gray entropy, BF-PSO.

1 Introduction

Image segmentation means to divide an image into two parts of the object and the background. It is the requirement of object identification and target tracking. Segmentation based on the threshold is widely used. But, the traditional methods only consider of the gray-level information, ignoring the spatial gradient information.

To this, Zhou et al.[1] propose a method based on gray-gradient co-occurrence matrix. The method makes use of the gray-level information and the gradient information at the same time. It enhances the quality of edges, but is sensitive to noise. Ma et al. [2] propose a method to construct the gray-gradient co-occurrence matrix based on wavelet transform, and search the best threshold via genetic algorithm. This method can inhibit noise to some extent. The reason is that the low-frequency coefficients contain general gray-level information, and the high-frequency coefficients contain the gradient information without noise.

The wavelet transform cannot take full advantage of the geometric feature of images, and it's not the optimal basis. The NSCT provides a complete shift-invariant and the function of multi-scale analysis. So this paper considers construct the gray-gradient co-occurrence matrix based on NSCT and search the best threshold via optimization algorithms. In recent years, swarm intelligence optimization methods become a hot topic. Particle swarm optimization (PSO) and bacterial foraging optimization (BFO) are two typical algorithms. PSO is faster but easy to fall into local

optimum; BFO algorithm shows the detailed search feature and global optimization ability, but requires more iteration. The bacterial foraging- particle swarm optimization (BF-PSO) algorithm is the combination of BFO and PSO, keeping the advantage and avoiding the shortage of the two algorithms[3].

To improve the segmentation quality and reduce the iteration number, this paper proposes a new method based on NSCT and BF-PSO. The method Constructs the gray-gradient co-occurrence based on NSCT, and searches the best threshold via BF-PSO (NSCT, BF-PSO).

2 Related Technologies

2.1 NSCT Technology

NSCT have got rapid development since it was proposed by Cunha et al. [4] in 2006. It is a shift-invariant version of the Contourlet Transform (CT). NSCT is based on Non Subsampled Pyramid Filter Banks (NSPFB) and Non Subsampled Directional Filter Banks(NSDFB). Fig. 1 shows the decomposition framework of NSCT. This filter banks can be achieved by using two-channel nonsubsampled 2-D filter banks.

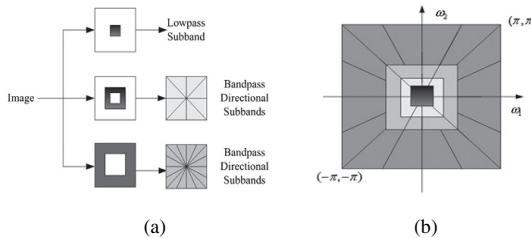


Fig. 1. Non Subsampled Contourlet Transform. (a) Non Subsampled Filter Bank structure that implements the NSCT. (b) The idealized frequency partitioning.

The NSPFB consists of several two-channel Non Subsampled Filter Banks, which is different from LP in CT. It is employed by the NSCT for its shift-invariance. The NSDFB are constructed by eliminating the down-sample in the DFB. To achieve multi-direction decomposition, the NSDFB is iteratively used. NSCT has been developed rapidly because of its multi-scale analysis property and shift-invariance.

2.2 Gray Entropy Model

Reference[1]gives the definition of gray-gradient co-occurrence matrix. This model can correctly reflect the distribution of gray-levels in both homogeneous region and texture region. In this paper, we construct a new co-occurrence matrix based on NSCT. The steps are given below.

1. Get a general fuzzy image F and a gradient image G of the original image I .

First, perform a 3-level NSCT decomposition on the origin image I . Then, reconstruct the low-frequency coefficient and it is regarded as image F . Reconstruct

the high-frequency coefficients and regard it as image G . F is a blurred version of I . In the contrary, G is a sharpened version, which has the numerical gradient of image I .

2. Normalize the images F and G to $[0, L-1]$.

$$F(m, n) = \text{round}\left(\frac{F(m, n) - \min(F(m, n))}{\max(F(m, n)) - \min(F(m, n))} \cdot (L-1)\right) \quad (1)$$

$$G(m, n) = \text{round}\left(\frac{G(m, n) - \min(G(m, n))}{\max(G(m, n)) - \min(G(m, n))} \cdot (L-1)\right) \quad (2)$$

Round is an operator to get an integer format; $\max(F(m, n))$, $\max(G(m, n))$, $\min(F(m, n))$ and $\min(G(m, n))$ represent the maximum and the minimum gray value in F and G , respectively; $L-1=255$ stands for the maximum value after normalized.

3. Construct a gray-gradient co-occurrence matrix Co .

Each element in Co c_{ij} is the number of pixel pairs, satisfying in $F(m, n)=i$, and $G(m, n)=j$. Then, the probability P_{ij} in Co can be expressed as

$$P_{ij} = \frac{c_{ij}}{\sum_i \sum_j c_{ij}} \quad (3)$$

During the course of segmentation, we suppose that the two thresholds in images F and G are s and t , which are grey numbers ranging from 0 to $L-1$. The position of (s, t) will divide Co into four quadrants, as shown in Fig. 3.

In A or D area, the gradient gray-level are small. A denotes the objects, while D denotes backgrounds in the image. On the contrary, in B and C the gradient gray-level are bigger. That means B and C are correspond to the edge and texture region.

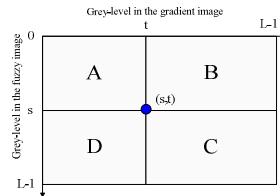


Fig. 2. Components of the gray-gradient co-occurrence matrix Co

Suppose that c_{ij} is an arbitrary element of B, representing the transferring number of i belonging to objects and j to the edges; p_{ij}^B is the probability of (i, j) transferring from objects to edges. Similarly, c_{ij} in C is the transferring number of i belonging to the background and j to edges; p_{ij}^C is the probability of (i, j) transferring from the background to edges. Below is the formula for calculate p_{ij}^B and p_{ij}^C .

$$p_{ij}^B = \frac{c_{ij}}{\sum_{i=0}^s \sum_{j=t+1}^{L-1} c_{ij}} \quad p_{ij}^C = \frac{c_{ij}}{\sum_{i=s+1}^{L-1} \sum_{j=t+1}^{L-1} c_{ij}} \quad (4)$$

The 2D conditional entropy is defined as below:

$$H(s, t) = \frac{1}{2}(H(E|O) + H(E|B)) = \frac{1}{2}\left[\left(-\sum_{i=0}^s \sum_{j=t+1}^{L-1} p_{ij}^B \log_2 p_{ij}^B\right) + \left(-\sum_{i=s+1}^{L-1} \sum_{j=t+1}^{L-1} p_{ij}^C \log_2 p_{ij}^C\right)\right] \quad (5)$$

When formula(5) used in image segmentation, we name the 2D conditional entropy containing a pair of grey number (s, t) as gray entropy(GE). And we can acquire the best position (s, t) corresponding to the maximum value of gray entropy.

2.3 BF-PSO Algorithm

1. Basic BFO algorithm

The BFO algorithm is proposed by K. M. Passino [5] in 2002. To imitate the foraging behavior of coli, three stages are required below:

(a) Chemotaxis step

This stage consists of two operations called swimming and tumbling. Swimming represents bacteria reach a new position by moving to a random direction. Bacteria keeps swimming to this direction until the nutrition concentration becomes lower or reaches the maximum number of swimming steps N_s . Either situation above occurs, perform tumbling operation. Bacterial populations search for high nutrition concentration regions in this way.

(b) Reproduction step

In this step, let N_{re} be the number of reproduction steps. S is the number of all bacteria and is divided into two parts,

$$S_r = S/2 \quad (6)$$

The ones with the lowest health dies and those with enough nutrients will be reproduced (split in two). Thus, the number of the bacteria is always S .

(c) Elimination and Dispersal Step

This step takes place after a certain number of reproduction processes. First, a probability of elimination and dispersal called P_{ed} is chosen for each bacterium. And then based on P_{ed} , it moves to another position in the environment.

2. Basic PSO algorithm

PSO algorithm is introduced by Eberhart and Kennedy[6] in 1995. When solving optimization problems using PSO, each solution is regarded as a particle in the problem space. The particles move iteratively through the d-dimension problem space to search the new solutions. In generation t , the position and velocity of the i -th particle are denoted by $x_i(t)$ and $v_i(t)$. Each particle remembers its own best position x_{pbest} . The best position among the swarm is named x_{gbest} . Particles move according to the formula below, updating their position and velocity.

$$v_i(t) = wv_i(t-1) + c_1r_1(t)(x_{pbest} - x_i(t)) + c_2r_2(t)(x_{gbest} - x_i(t)) \quad (7)$$

$$x_i(t) = x_i(t-1) + v_i(t) \quad (8)$$

In which, w is nonnegative inertia weight; c_1, c_2 are nonnegative learning factors; r_1, r_2 are random numbers in $[0,1]$.

3. BF-PSO algorithm

The BF-PSO algorithm combines the two algorithms together. This combination aims to reduce the number of iteration in BFO, and avoid trapping in local optimal points which usually happens in PSO. The BF-PSO algorithm flow is below:

- (a) Initialize parameters.
- (b) Compute the Fitness value
Compute the fitness value of each bacterium.
- (c) Iterative optimization
There are 3 loops:
The inner-loop is to update the position of the bacteria;
The middle-loop is the reproduction step;
The outer-loop is to perform the elimination and dispersal operation.
- (d) Output optimal solution

3 Image Segmentation Method Based on NSCT and BF-PSO

The central idea of our method is to employ the gray entropy based on NSCT to be the fitness function of the BF-PSO algorithm. The best threshold (s, t) is achieved when the entropy reaches its maximum value. Then, take s as the segmentation threshold, and complete the segmentation. The main procedure is illustrated in Fig. 4.

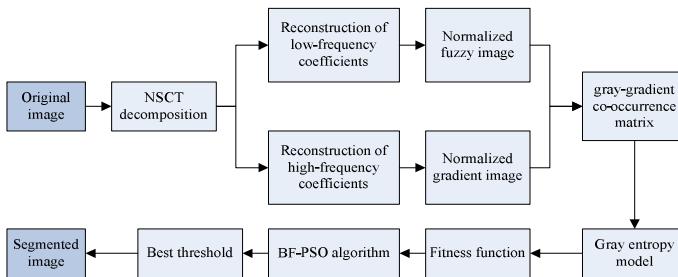


Fig. 3. Procedure of image segmentation method based on NSCT and BF-PSO

- (a) Perform NSCT decomposition on original image I ;
- (b) Reconstruct the low-frequency and the high-frequency coefficients respectively. Normalize the matrixes to get the general fuzzy image F and gradient image G .
- (c) Construct the co-occurrence matrix and create the gray entropy model.
- (d) Design the fitness function of BF-PSO algorithm based on the gray entropy model. Calculate the best threshold (s, t) via BF-PSO algorithm.
- (e) Take s as the threshold for segmentation, perform segmentation operation on the original image I . For the pixels whose gray value is lower than s , Assign them a value of zero. For the ones with higher gray value, assign them a value of 255.

4 Experimental Results and Analysis

To test the effectiveness of our method, make a comparison between it and three other methods. The other methods are mentioned in paper[7,8], and we improved them to a certain extent to make them more comparable. They are listed below:

- (a) image segmentation method based on NSCT and BFO (NSCT, BFO);
- (b) image segmentation method based on NSCT and PSO (NSCT, PSO);
- (c) image segmentation method based on Wavelet and BF-PSO (Wavelet, BF-PSO).

Take Gray-level Contrast(GC) and Probability of Error(PE) as evaluation standard. GC is calculated according to formula (9).

$$GC = \frac{|f_1 - f_2|}{f_1 + f_2} \quad (9)$$

In the formula above, f_1, f_2 represents the average gray level of the object region and the background region, respectively.

When calculating PE, take an artificial segmentation result as the reference.

$$PE = \frac{P_{error}}{P_{object}} \quad (10)$$

P_{error} is the number of error pixels. Error pixels contain the pixels mistaken for belonging to the background region which belong to the object region in fact and the pixels mistaken for belonging to the object region which belong to the background region. P_{object} is the number of pixels in the object region.

The threshold is shown in each experiment. It is the value of s when the gray entropy reaches its maximum in the gray entropy model.

Experiment 1

In this experiment, use a simulated image ‘eight’ with size 242×308 as the test image. The segmentation results are shown in Fig. 5.

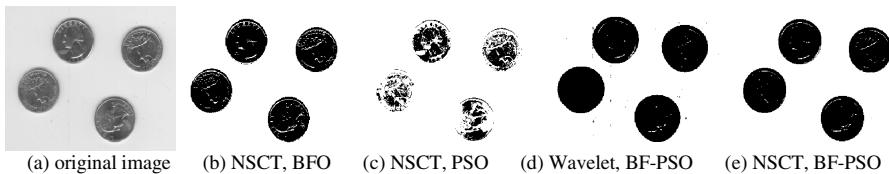


Fig. 4. Image ‘eight’ and its segmentation results

Obviously, the segmentation performance of the method proposed in this paper outperforms the others. In Fig. 5., the results of (b) and (c) are insufficient segmentations, while the result of (d) is over segmentation. In deed, our method needs less iteration as well. The evaluation parameters are shown in Table 1 and the comparison of iteration numbers are shown in Fig. 6.

Table 1. Comparison of different methods in Experiment 1

Segmentation methods	s	GC	PE
NSCT, BFO	134	0.3866	0.0795
NSCT, PSO	94	0.4213	0.6008
Wavelet, BF-PSO	214	0.3156	0.2255
NSCT, BF-PSO	161	0.3705	0.0209

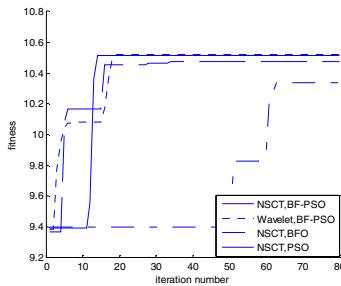


Fig. 5. Iteration numbers of different methods in Experiment 1

Experiment 2

In this experiment, take a noisy image ‘chromosome’ with size 264×338 as the test image. The results of different methods are shown in Fig. 7.

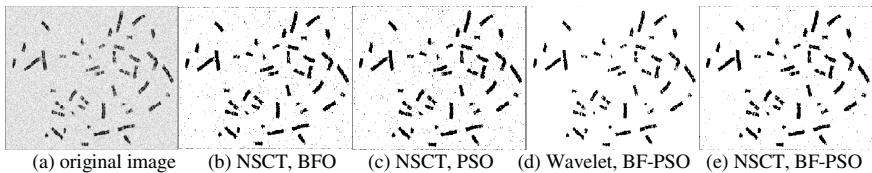


Fig. 6. Image ‘chromosome’ and its segmentation results

It is easy to observe that in (b) and (c) there are quite a few noise points. The result in (d) appeared object missing phenomenon. The segmentation result in (e) has the fewer noise points avoiding object missing. Table 2 and Fig. 8 are showing the comparison of different methods.

Table 2. Comparison of different methods in Experiment 2

Segmentation methods	S	GC	PE
NSCT, BFO	182	0.3668	0.1978
NSCT, PSO	183	0.3618	0.2166
Wavelet, BF-PSO	144	0.4740	0.1300
NSCT, BF-PSO	175	0.3981	0.0903

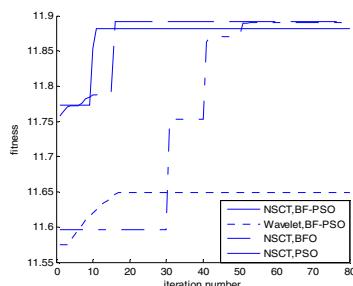


Fig. 7. Iteration numbers of different methods in Experiment 2

Experiment 3

A test image of Berkeley Segmentation Dataset is used in this experiment. The results of different methods are shown below.

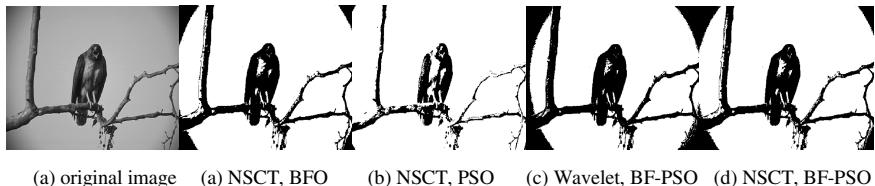


Fig. 8. Image of Berkeley Dataset and its segmentation results

As seen in Fig. 8, the result in (a) and (b) has lost some of the objects. The result in (c) has less detail information. The result of our method has no target losing, and it has rich detail information. The objective evaluation index is shown in Table 3. The iteration numbers are shown in Fig. 9.

Table 3. Comparison of different methods in Experiment 3

Segmentation methods	s	GC	PE
NSCT, BFO	126	0.4180	0.0844
NSCT, PSO	74	0.5732	0.3987
Wavelet, BF-PSO	148	0.3300	0.4642
NSCT, BF-PSO	114	0.4578	0.0516

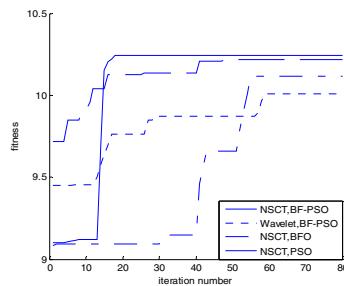


Fig. 9. Iteration numbers of different methods in Experiment 3

5 Conclusions

A new image segmentation method based on NSCT and BF-PSO algorithm is proposed. Experiments on a simulated image, a noisy image and a test image of Berkeley Segmentation Dataset show that the method is efficient. The results also show that the method proposed in this paper can make better use of the edge information, and need less iteration.

Acknowledgment. The research is supported by “the Fundamental Research Funds for the Central Universities” under grant number 13XS35.

References

1. Delong, Z., Quan, P., Hongcai, Z.: Maximum Entropy Thresholding Algorithm. *Journal of Software* 12(9), 1420–1422 (2001)
2. Ma, M., Lu, Y., Zhang, Y., He, X.: A Fast SAR Image Segmentation Algorithm Based on Two Dimension Gray Entropy Model. *Journal of Xidian University (Natural Science Edition)* 36(6), 1114–1119 (2009)
3. Korani, W.M., Dorrah, H.T., Emara, H.M.: Bacterial Foraging Oriented by Particle Swarm Optimization Strategy for PID Tuning. In: *IEEE International Symposium on Digital Object Identifier, Computational Intelligence in Robotics and Automation(CIRA)*, pp. 445–450 (2009)
4. Da Cunha Arthur, L., Zhou, J., Do, M.N.: The nonsubsampled contourlet transform: Theory, design, and applications. *IEEE Transactions on Image Processing* 15(10), 3089–3101 (2006)
5. Passino, K.M.: Biomimicry of Bacterial Foraging for Distributed Optimization and Control. *IEEE Control Systems Magazine* 22(3), 52–67 (2002)
6. Kennedy, Eberhart, R.C.: Particle Swarm Optimization. In: *Proc. of the IEEE Int. Conf. on Neural Networks*, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
7. Ma, M., Liang, J., Guo, M.: SAR Image Shreshold Segmentation Based on Bacterial Foraging Algorithm. *Journal of Xidian University (Natural Science Edition)* 38(6), 152–158 (2011)
8. Li, L.: Research on the Technology of Video Text Information Extraction, pp. 71–76. Harbin Engineering University (2012)

Non-linear Feature Fusion Based on Polynomial Correlation Filter for Face Recognition

Dong Yan, Yuanyuan Shen, Yan Yan^{*}, and Hanzi Wang

School of Information Science and Technology, Xiamen University, China

{yd.sunday, shenyuanyuan1989}@gmail.com,

{yanyan, hanzi.wang}@xmu.edu.cn

Abstract. Face recognition is an active research area due to its wide range of practical applications. Efficient and discriminative facial feature is a crucial issue for face recognition. Most existing methods use one type of features but we show that robust face recognition requires different kinds of feature information to be taken into account. Traditional feature fusion methods are based on the linear combination. In this study, we propose a novel and effective fusion method (called NF-PCF), which uses polynomial correlation filter (PCF) to non-linearly fuse different types of features for robust face recognition. Experimental results on two popular face databases, including Yale and PIE, show the promising results obtained by the proposed method.

Keywords: Face recognition, feature fusion, non-linear fusion, correlation filter.

1 Introduction

Technologies of face recognition have been significantly developed in recent decades, but it is still challenged by some tough problems such as facial expression, large variations in illumination, pose, aging, and partial occlusions [1]. These challenges make robust face recognition become a hard problem, which requires using effective facial features to handle drastic facial appearance variations of one person and great similarities between different persons.

A large number of facial features have been proposed, including the global features (such as gray scale [2]) and the local features (such as SIFT [3], HOG [4] and LBP [5]). Each type of features characterizes facial information in its own way. For instance, the gray scale is a global feature that describes a facial appearance from the visual perspective. SIFT [3] is widely used in object matching, which can locate key points in images at different scales. However, the features extracted by SIFT may not be enough for face recognition since it mainly concentrates on key facial points. To some degree, the disadvantages of SIFT have inspired the development of Histogram of Oriented Gradients (HOG) [4], which has been proved effective in pedestrian detection. HOG uses the histogram of gradients to represent the local appearance and

^{*} Corresponding author.

shape of a subject, thus making it possible to extract effective features for face recognition. LBP [5] is another effective feature for face recognition which encodes fine details of facial texture.

To the best of our knowledge, there is no single feature which can handle all the situations of facial appearance changes. Therefore, finding and combining complementary features that are robust to facial appearance changes has become a crucial problem for robust face recognition. Global features can cope well with images under uniform illumination changes, but they are sensitive to variations in facial expression, pose. On the other hand, local features are more robust to local appearance changes caused by expression, pose, etc. In this study, inspired by the complementary advantages of global feature and local feature, we propose a novel and effective feature fusion method, called NF-PCF, which uses polynomial correlation filter (PCF) to non-linearly fuse two different types of features (i.e., gray scale and HOG) for robust face recognition. Compared with the traditional feature fusion methods, NF-PCF tries to fuse global feature and local feature in a non-linear way. In addition, NF-PCF is jointly optimized by emphasizing the combined outputs of different features.

Information fusion in the pattern recognition field can be roughly classified into two categories: feature-level [6,7] fusion and decision-level fusion [8,9]. Conventional feature-level fusion methods concentrate different features into a single feature, while the decision-level fusion methods integrate the outputs of several classifiers to make the final decision [10]. Chowdhury et al. [6] proposed to extract the discriminant feature vector by concatenating local discriminant features obtained by 2D-LDA method. Liu et al. [7] directly combined the color, local spatial and global frequency information to constitute the final feature vector. Jain et al. [8] combined multiple classification scores by using a weighted sum rule.

In this study, we use PCF to effectively fuse features rather than implementing a direct concentration in the feature-level. We show that the performance of a face recognition method can be significantly improved by NF-PCF. Experimental results on two well-known face databases (i.e., Yale and PIE face databases) show that NF-PCF has achieved the promising results. In summary, the main contribution of our work is that we exploit both the merits of global and local feature by using PCF to perform effective feature fusion. Moreover, we show that non-linear fusion can be beneficial to improve the performance in face recognition.

The rest of this paper is organized as follows. Section 2 describes the concept of the correlation filter used in this paper. Section 3 shows the details of the proposed method which uses the non-linear feature fusion based on the polynomial correlation filter (i.e., NF-PCF). Section 4 presents and analyzes the experimental results. Section 5 makes the conclusions.

2 Related Work

Correlation filters have been successfully applied to deal with facial appearance variations due to their attractive properties such as shift-invariance, graceful degradation,

and closed-form solutions [11,12]. Many types of correlation filters [13,14] have been proposed in recent decades.

As our method is based on optimal tradeoff filter (OTF) [14], we now briefly review the design of OTF. Suppose we have N training images and the size of each image is $r \times c$. The basic steps of OTF are summarized as follows. First, each image is vectorized as a K -dimensional ($K = r \times c$) column vector. Thus, the training set can be represented as a matrix \mathbf{Q} ($\mathbf{Q} \in R^{K \times N}$). Then, each column is transformed to the frequency domain using the fast Fourier transform (FFT). Finally, the solution of OTF is obtained based on the frequency domain, which can be written as:

$$\mathbf{h} = \mathbf{T}^{-1} \mathbf{M} (\mathbf{M}^* \mathbf{T}^{-1} \mathbf{M})^{-1} \mathbf{u} \quad (1)$$

where $\mathbf{u} = [u_1, u_2, \dots, u_N]$ is a $N \times 1$ vector and u_t denotes the correlation peak amplitude of the t -th training image. u_t is equal to 1 for the authentic images and 0 for the imposter images. \mathbf{M} is the 1D Fourier transforms of matrix \mathbf{Q} and ‘ $+$ ’denotes the conjugate transpose. $\mathbf{T} = \alpha \mathbf{D} + (1 - \alpha) \mathbf{C}$, where \mathbf{D} is a diagonal matrix whose diagonal values are the average power spectrum of the N training images; \mathbf{C} is a diagonal matrix (where the input noise power spectral density values are used as its diagonal values). α is a trade-off parameter.

3 Non-linear Feature Fusion Based on Polynomial Correlation Filter (NF-PCF)

This section describes the details of our proposed NF-PCF in details. The flow diagram of the proposed NF-PCF is illustrated in Fig. 1. It contains four steps: feature extraction; dimension reduction by CFA [12], where the filters are designed based on PCF; summation of the outputs; classification based on the Nearest Neighbor (NN) method. Next, we explain the dimension reduction by CFA in Sections 3.1. The design of PCF is described in Section 3.2.

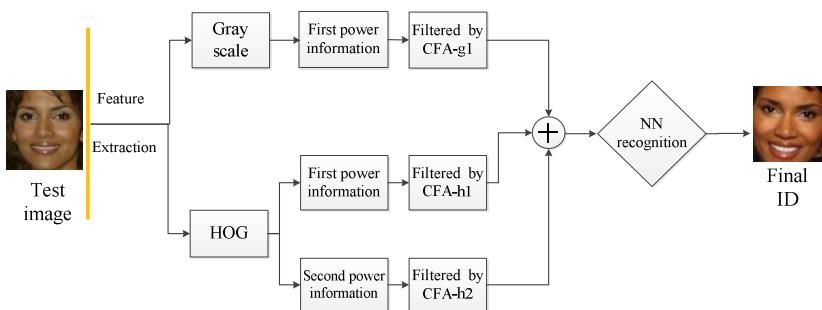


Fig. 1. The flow diagram of the proposed NF-PCF

3.1 Dimension Reduction by CFA

Traditional fusion methods use PCA [2] or other dimension reduction methods [15-18] to transform different types of features to the same low-dimensional features and then fuse these low-dimensional features [19]. However, such strategy may lose useful information when enforcing the same reduced dimension for each type of feature. On the other way, the projection matrix obtained by CFA is fixed (which is the same as the number of training classes), most information can be preserved after the projection by CFA. Therefore, in this study, we use CFA to reduce different features to the same dimension.

In the proposed method, three different OTFs (one for gray scale, two for HOG with the first and second power) are respectively designed in PCF (see Section 3.2). Suppose that we have N training images which are divided into L classes. Thus, we obtain the gray scale matrix ($\mathbf{X}_{train_g} \in R^{d_1 \cdot N}$, where d_1 is the size of the gray scale feature) and two HOG matrices ($\mathbf{X}_{train_h}^1 \in R^{d_2 \cdot N}$ represents the 1st power HOG feature, $\mathbf{X}_{train_h}^2 \in R^{d_2 \cdot N}$ represents the 2nd power HOG feature, where d_2 is the size of the HOG feature) for all the face images. Then, PCF is designed based on three feature matrices for each class, where three OTFs are simultaneously designed. Finally, a projection matrix is computed by combining the corresponding OTFs of all the classes for each type of feature.

On the other hand, the gray scale feature and HOG feature extracted from a test image are represented as $\mathbf{x}_{test_g} \in R^{d_1 \cdot 1}$, $\mathbf{x}_{test_h}^1 \in R^{d_2 \cdot 1}$, and $\mathbf{x}_{test_h}^2 \in R^{d_2 \cdot 1}$ respectively. The dimension reduction by CFA can be expressed as: $\mathbf{y}_{test_g} = \mathbf{P}_{g1}^T \mathbf{x}_{test_g}$, $\mathbf{y}_{test_h1} = \mathbf{P}_{h1}^T \mathbf{x}_{test_h}^1$ and $\mathbf{y}_{test_h2} = \mathbf{P}_{h2}^T \mathbf{x}_{test_h}^2$, where $\mathbf{P}_{g1} \in R^{d_1 \cdot L}$, $\mathbf{P}_{h1} \in R^{d_2 \cdot L}$ and $\mathbf{P}_{h2} \in R^{d_2 \cdot L}$ represent the projection matrices obtained by the gray scale and HOG features corresponding to the first and second power. \mathbf{x}_{test_g} and $\mathbf{x}_{test_h}^i$ ($i=1,2$) are respectively the input gray scale feature and the HOG feature with the power i . $\mathbf{y}_{test_g} \in R^{L \cdot 1}$ and $\mathbf{y}_{test_hi} \in R^{L \cdot 1}$ ($i=1,2$) indicate the projected features. It is worth to point out that after CFA, different features have a same dimension, thus making it possible to implement a weighted fusion operation.

After projection, we represent different types of features by the corresponding low-dimensional features, and these features make the fusion more effective. We should point out that, for the gray scale feature, only the first power is considered. This is due to the fact that the non-linear transform of the global feature can result in losing useful information for face recognition. However, the non-linear transform of the local feature, which captures the local relationship of the facial shape and texture, is helpful to increase the discriminability [20]. After the feature extraction, the nearest neighbor method is finally used for classification.

3.2 The Design of PCF

Now we explain the design of Polynomial Correlation Filters (PCF). PCF is an advanced composite correlation filter. The main difference between PCF and the traditional correlation filters is that the output of PCF is a non-linear function of the inputs, as shown in Fig. 2.

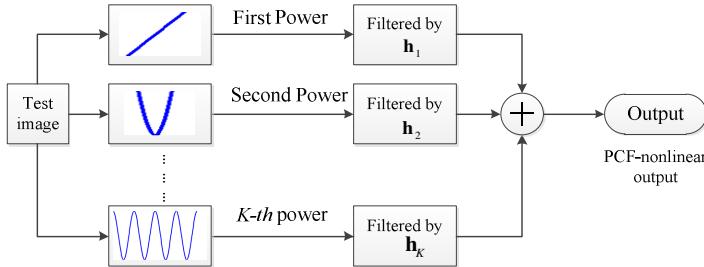


Fig. 2. Illustration of PCF

The traditional PCF combines multiple filters based on different power information to form a composite classifier. The objective function of the traditional PCF can be written as:

$$P(\mathbf{h}_m, \mathbf{x}) = \mathbf{h}_{m1}\mathbf{x}^1 + \cdots + \mathbf{h}_{mi}\mathbf{x}^i + \cdots + \mathbf{h}_{mK}\mathbf{x}^K \quad (2)$$

where \mathbf{x}^i is a column vector representing a facial feature and each of its elements is raised to the power i for the image \mathbf{x} . $\mathbf{h}_m = [\mathbf{h}_{m1}; \dots; \mathbf{h}_{mK}]$ is the correlation filter array corresponding to K different facial features designed for class m . And $P(\mathbf{h}_m, \mathbf{x})$ is a value which represents the polynomial correlation output of \mathbf{x} .

OTF is used as the single correlation filter in PCF. $P(\mathbf{h}_m, \mathbf{x})$ indicates the similarity between the image \mathbf{x}_n with the gallery class m . K is the highest power in PCF. A higher K value can further improve the performance, but the computational complexity is also increased. In our work, K is set to 2.

Compared with the traditional PCF which uses only one feature (using different powers) as the inputs, we propose to use and fuse two different features for PCF. The design flow of PCF is described as follows. First, the training set is used to build three OTFs. The feature sets are transformed according to the order of power. Based on Eq. (2), the proposed PCF with feature fusion can be written as:

$$P(\mathbf{h}_m, \mathbf{x}) = \mathbf{h}_{mG}\mathbf{x}_G^1 + \mathbf{h}_{mH1}\mathbf{x}_H^1 + \mathbf{h}_{mH2}\mathbf{x}_H^2 \quad (3)$$

where \mathbf{h}_{mG} and \mathbf{h}_{mHi} ($i=1, 2$) respectively represent the correlation filters designed for the gray scale feature and HOG feature with the power i for the m -th class. \mathbf{x}_G^i ($i=1, 2$) and \mathbf{x}_H^i ($i=1, 2$) represent the gray scale feature and the HOG feature for the input feature with the power i .

After some calculations, Eq. (3) can be reformulated as:

$$P(\mathbf{h}_m, \mathbf{x}) = \mathbf{h}_m^T \mathbf{x} \quad (4)$$

where $\mathbf{h}_m = [\mathbf{h}_{mG}; \mathbf{h}_{mH1}; \mathbf{h}_{mH2}]$ and $\mathbf{x} = [\mathbf{x}_G^1; \mathbf{x}_H^1; \mathbf{x}_H^2]$.

Therefore, PCF has a closed-form solution which is the same as Eq. (1). However, compared with OTF, PCF uses a non-linear function for the inputs. In addition, three different OTFs are designed simultaneously by solving PCF.

4 Experiments

This section gives the experimental results. In Section 4.1, experimental settings are given. Results and analysis are shown in Section 4.2

4.1 Experimental Settings

In this section, experimental results on two popular face databases, including Yale [20], PIE [21], are evaluated to illustrate the effectiveness of the proposed NF-PCF. For the Yale database, all the images (containing 15 persons with 165 images) are used. For the PIE database, we choose 1,836 images with 68 classes (i.e., 27 images for each person), where the images are captured with severe pose and illumination variations.

All the face images are cropped and normalized to the size of 64×64. For all the databases, we randomly choose 30% of the data as the training set and the rest is used as the test set. The experiments are repeated 20 times. The average recognition rates are reported. We compare the proposed NF-PCF with several state-of-the-art recognition methods, including PCA [2], Fisherface [17], CFA [12], PCA+CFA [22], SRC [23], LPP [24], PCF (with only one feature, $K = 2$) [19]. Our NF-PCF is also compared with C-PCF, where different features are directly concentrated as a single feature to design the traditional PCF. In Table 1, $g=1$ (or $g=2$) indicates that NF-PCF is designed by using the 1st (or 2nd) power of the gray scale feature; $h=1$ (or 2) represents it is designed by using the HOG feature with 1st power (or 2nd power). For simplicity, we call NF-PCF (with $g=1$, and $h=1,2$) as NF-PCF1; we call NF-PCF (with $g=2$, and $h=1,2$) as NF-PCF2.

Table 1 shows the average recognition rates obtained by the competing methods on the two databases. Results obtained on the Yale database show that the results obtained by PCF, C-PCF and NF-PCF are better than CFA. This result validates that the second power information of a feature can be helpful to improve the performance of a face recognition method. Furthermore, NF-PCF1 obtains better recognition rate compared with C-PCF, which shows that the proposed non-linear fusion is more effective than the direct concentration strategy. NF-PCF1 has achieved the best results on the PIE database, which shows that NF-PCF1 has the better generalization ability than the other competing methods. In this case, NF-PCF1 outperforms CFA about 11% and obtains the highest recognition rates (47.65%) compared with other methods. Therefore, non-linear fusion is beneficial for improving the final recognition rate of the proposed method.

Table 1. The average recognition rates obtained by the competing methods on the Yale and PIE databases

Method	Yale database		PIE database	
	HOG	Gray scale	HOG	Gray scale
PCA	56.67% \pm 2.00	60.00% \pm 1.50	25.74% \pm 1.15	26.40% \pm 1.09
Fisherface	67.50% \pm 1.55	63.33% \pm 2.58	29.04% \pm 1.36	30.88% \pm 1.06
CFA	70.00% \pm 1.67	81.67% \pm 0.67	29.56% \pm 0.89	36.00% \pm 3.68
PCA+CFA	68.33% \pm 1.83	81.67% \pm 2.17	29.56% \pm 1.27	39.71% \pm 0.97
SRC	75.00% \pm 1.64	82.50% \pm 1.28	36.40% \pm 1.60	45.59% \pm 1.66
LPP	65.83% \pm 2.23	70.83% \pm 2.45	27.35% \pm 2.36	21.54% \pm 2.66
PCF	70.83% \pm 1.25	82.67% \pm 2.33	28.53% \pm 2.07	34.19% \pm 1.96
C-PCF ($g=1, h=1, 2$)	84.17% \pm 2.58		40.96 % \pm 2.17	
C-PCF ($g=2, h=1, 2$)	73.33% \pm 0.75		29.63% \pm 1.75	
NF-PCF1	86.67% \pm 2.17		47.65% \pm 1.18	
NF-PCF2	75.00% \pm 1.92		34.04% \pm 1.97	

5 Conclusions

In this paper, we propose an effective feature fusion method (called NF-PCF) to non-linearly fuse two complementary features (i.e., gray scale and HOG features) for improving face recognition performance. An effective polynomial correlation filter is employed to combine the gray scale and HOG features. Experimental results on two popular databases with variations in illumination, expression and pose demonstrate that the proposed NF-PCF can obtain promising recognition performance and outperform several competing methods for most databases.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 61201359 and 61170179, by the Natural Science Foundation of Fujian Province of China under Grant 2012J05126, by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20110121110033.

References

1. Zhao, W.Y., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. *ACM Comput. Surv.* 35(4), 399–458 (2003)
2. Turk, M.: Pentland: Eigenfaces for Recognition. *J. Cogn. Neurosci.* 3(1), 71–86 (1991)
3. Lowe, D.: Distinctive Image Features from Scale-Invariant Key points. *Int. J. Comput. Vision* 60(2), 91–110 (2004)
4. Navneet, D., Bill, T.: Histograms of Oriented Gradients for Human Detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)

5. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(12), 2037–2041 (2006)
6. Chowdhury, S., Sing, J., Basu, D., Nasipuri, M.: Face Recognition by Fusing Local and Global Discriminant Features. In: Proc. Second International Conference on Emerging Applications of Information Technology, pp. 102–105 (2011)
7. Liu, Z.M., Liu, C.J.: Fusion of Color, Local Spatial and Global Frequency Information for Face Recognition. *Pattern Recognition* 43(8), 2882–2890 (2010)
8. Jain, A., Nandakumar, K., Ross, A.: Score Normalization in Multimodal Biometric Systems. *Pattern Recognition* 38(12), 2270–2285 (2005)
9. Cyran, K.A., Kawulok, J., Kawulok, M., Stawarz, M., Michalak, M., Pietrowska, M., Widlak, P., Polanska, J.: Support Vector Machines in Biomedical and Biometrical Applications. In: Proc. Emerging Paradigms in Machine Learning, pp. 379–417 (2013)
10. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(3), 226–239 (1998)
11. Vijaya Kumar, B.V.K., Savvides, M., Xie, C.: Correlation Pattern Recognition for Face Recognition. *Proc. IEEE* 94(11), 1963–1976 (2006)
12. Mahalanobis, A., Vijaya Kumar, B.V.K., Casasent, D.: Minimum Average Correlation Energy Filters. *Appl. Opt.* 26(17), 3630–3633 (1987)
13. Vijaya Kumar, B.V.K.: Minimum Variance Synthetic Discriminant Functions. *J. Opt. Soc. Amer.* 3(10), 1579–1584 (1986)
14. Refregier, P.: Filter Design for Optical Pattern Recognition: Multi-Criteria Optimization Approach. *Opt. Lett.* 15(15), 854–856 (1990)
15. Deniz, O., Bueno, G., Salido, J., De la Torre, F.: Face Recognition Using Histograms of Oriented Gradients. *Pattern Recognition Letters* 32(12), 1598–1603 (2011)
16. Junior, O., David, D., Goncalves, V., Nunes, U.: Trainable Classifier-Fusion Schemes: an Application to Pedestrian Detection. In: Proc. ITSC, pp. 1–6 (2009)
17. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces Versus Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 711–720 (1997)
18. Liu, C., Wechsler, H.: A Shape and Texture Based Enhanced Fisher Classifier for Face Recognition. *IEEE Trans. Image Processing*. 10(4), 598–608 (2001)
19. Mahalanobis, A., Vijaya Kumar, B.V.K.: Polynomial Filters for Higher Order Correlation and Multi-input Information Fusion. In: Proc. SPIE, pp. 221–231 (1997)
20. Georgiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(6), 643–660 (2001)
21. Sim, T., Baker, S., Bsat, M.: The CMU Pose Illumination and Expression (PIE) Database of Human Faces. In: Proc. AFGR, pp. 46–51 (2002)
22. Yan, Y., Zhang, Y.-J.: 1D Correlation Filter Based Class-dependence Feature Analysis for Face Recognition. *Pattern Recognition* 41(12), 3834–3841 (2008)
23. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S., Ma, Y.: Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(2), 210–227 (2008)
24. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face Recognition Using Laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(3), 328–340 (2005)

Robust Head Pose Estimation with a New Principal Optimal Tradeoff Filter

Dong Yan, Yan Yan, and Hanzi Wang

School of Information Science and Technology, Xiamen University, China
yd.sunday@gmail.com, {yanyan, hanzi.wang}@xmu.edu.cn

Abstract. One of the challenging problems encountered by face recognition is the difficulty in tackling pose variations, where fast and reliable head pose estimation is an essential step. In this paper, based on the correlation filter technique, a novel feature extraction framework, i.e., directional correlation filter set (DCFS), is developed for robust head pose estimation. In this framework, a principal optimal tradeoff filter (called POTF) is designed in the feature subspace obtained according to principal component analysis (PCA). Compared with the traditional methods that rely on the exact localization of facial feature points, our proposed method exploits the 1D frequency domain of the training data by using the correlation filter technique, which can capture the high-order statistics of a face for effective head pose estimation. Experimental results on several public face databases with large pose variations, including PIE, HPI, and UMIST, show the promising performance obtained by the proposed method on head pose estimation.

Keywords: Head pose estimation, feature extraction, correlation filter.

1 Introduction

Face recognition (FR) has attracted much attention in recent decades. A large number of FR methods have been developed towards an automatic and robust FR system [1]. However, there are still several issues unsolved, such as FR under occlusions and variations in illumination, aging, pose, and so on [1]. Among these issues, FR under pose variations is one of the most challenging problems [2-7]. The performance of a face recognition system may drop significantly when large pose variations are present. To deal with pose variations in FR, a fast and reliable head pose estimation method is required.

Head pose estimation [2] can be roughly classified into two categories: model-based methods [8, 9] and appearance-based methods [4, 10, 11]. Model-based methods construct 3D models of human heads, and thus require the labeling of facial key points, which needs a large number of samples during training. On the other hand, appearance-based methods extract facial feature from a whole image. Most appearance-based methods depend on pose-invariant local features or the localization of key facial feature points [4]. For instance, Ho and Chellappa [5] proposed to use the dense SIFT [6] features for head pose estimation. Dantone et al. [4] demonstrated the

benefits of conditional regression forests by modeling the appearance and location of facial feature points that is conditionally dependent to the head pose. Kim et al. [12] estimated head pose based on a part-based face matching algorithm.

Among decades of development, many types of correlation filters [13] have been proposed. For example, Mahalanobis et al. [14] proposed the minimum average correlation energy (MACE) filter. Kumar [15] proposed the minimum variance synthetic discriminant function (MVSDF) filter. The optimal tradeoff filter (OTF) in [16] combines the MACE filter and the MVSDF filter to produce sharp correlation peaks and suppress noise, thus making it effective for pattern recognition. Correlation filter has been shown to be effective on the task of face recognition [7] due to their desirable properties, such as graceful degradation, shift-invariance and closed-form solutions. In fact, correlation filter is not only employed in face recognition, but also applied in other pattern recognition fields, such as fingerprint verification [17], palmprint identification [18].

In this paper, we propose an effective feature extraction framework, i.e., directional correlation filter set (DCFS), for robust head pose estimation based on the correlation filter technique. In this framework, a principal optimal tradeoff filter (called POTF) is designed on the principal component analysis (PCA) feature subspace in the 1D frequency domain, and we use the correlation outputs of different correlation filters to estimate a head pose. One advantage of the propose method is that it does not require the localization of key facial feature points used in traditional methods. Experimental results show that the 3D feature is more effective than traditional features for head pose estimation.

The main contributions of this paper are two-fold:

- An effective feature extraction framework called DCFS is proposed for head pose estimation. For each head pose, each correlation filter in DCFS is trained and its origin peak value is used to constitute the final pose features. In DCFS, we consider head pose estimation with 3 different yaw poses (i.e. left-profile, frontal, and right-profile). Hence, only one 3D feature is generated for feature extraction.
- We develop a principal optimal tradeoff filter (called POTF) for the 1D-OTF. In this paper we use PCA to preserve the dominant information and design the 1D-OTF in the feature subspace, which can significantly reduce the computational cost and improve the robustness to variations caused by illumination, expression, etc.

The paper is organized as follows. Section 2 explains the methodology of the proposed method in details. Section 3 shows experimental results on head pose estimation. Finally, we make some concluding remarks in Section 4.

2 Methodology

In this section, we will explain our methodology in details. In Section 2.1, a novel feature extraction framework for head pose estimation is presented. The design of POTF in this framework is further given in Section 2.2.

2.1 Feature Extraction Framework for Head Pose Estimation

The proposed feature extraction framework (i.e., DCFS) for head pose estimation is illustrated in Fig. 1. An input face image is first represented as a high-dimensional feature. After that, principal component analysis (PCA) is used to perform dimension reduction so that the prominent information of the face is preserved. Moreover, our directional filter bank consists of three correlation filters corresponding to left-profile, frontal, and right-profile, respectively, which is respectively used to correlate with the low-dimensional PCA features. Here, DCFS only concerns the pose information and ignores the person identity information. More specifically, each correlation filter in the DCFS tries to discriminate one specific pose from all the other poses. Finally, a direction feature is obtained for head pose estimation. In this paper, the pose of a head is estimated by using the simple nearest neighbor classifier, which is based on the Euclidean distance of the direction features. It is worth noting that a directional filter bank [19] has been proposed for the directional decomposition of images. However, it only extracts different directional information in the image domain. DCFS considers the PCA feature domain which can effectively reduce the noise and preserve the dominant information of images.

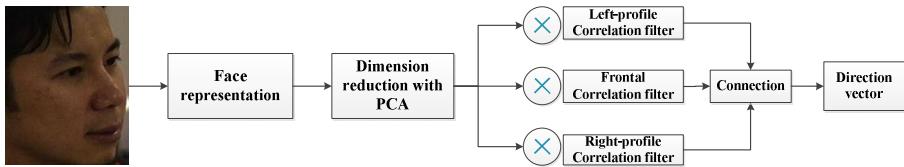


Fig. 1. Feature extraction framework for head pose estimation

2.2 Design of POTF

In the DCFS framework, the principal optimal tradeoff filter (called POTF) is a key constitution. In this section, we mainly focus on the design of POTF. The Optimal Tradeoff Filter (OTF) has been successfully applied in face recognition [7], but its computational complexity is high. Besides, OTF is relatively sensitive to variations in illumination and facial expressions due to the fact that the filter is designed in the original 2D image space. Therefore, we propose to design the correlation filters in 1D frequency domain on a feature subspace obtained by PCA.

The main steps of designing POTF are as follows. First, the PCA is first used to preserve the dominant information and remove the noise in the image. Second, we design the 1D-OTF in the PCA feature subspace. Therefore, the advantages of POTF are that the computational cost can be significantly reduced and the variations caused by illumination and pose can be alleviated.

Different from the traditional OTF which is based on the 2D image space, we proposed to design 1D-OTF in the low-dimensional PCA subspace. A directional correlation filter set (DCFS) consisting of three 1D-OTF correlation filters (each filter corresponding to a specific pose tries to distinguish one pose from the other two) is obtained.

1D-OTF is derived by combining the 1D-MACE (Minimal Average Correlation Energy) filter and the 1D-MVSDF (Minimal Variance Synthetic Discriminant Function) filter. The objective of the 1D-MACE filter is to minimize the average correlation energy (ACE), which can be formulated as

$$\begin{aligned} \min \frac{1}{N} \sum_{t=1}^N \sum_{p=0}^{m-1} |g_t(p)|^2 &= \min \frac{1}{Nm} \sum_{t=1}^N \sum_{p=0}^m |G_t(p)|^2 \\ &= \min \frac{1}{Nm} \sum_{t=1}^N \sum_{p=0}^m |F(p)|^2 |Y_t(p)|^2 \\ &= \min_{\mathbf{F}} \frac{1}{m} \mathbf{F}^T \mathbf{Q} \mathbf{F} \end{aligned} \quad (1)$$

where $G_t(p)$, $F(p)$ and $Y_t(p)$ are the 1D Fourier transforms of the output g_t , the correlation filter f and the input y_t , respectively; \mathbf{F} is the vector version of $F(p)$; ‘+’ means the conjugate transpose; \mathbf{Q} is a diagonal matrix whose diagonal entries are the average power spectrum of all N features.

The origin value of the correlation output is $g_t(0)=\mathbf{Y}_t^T \mathbf{F}$. The constraints of the MACE filter are that the values of the outputs at the origin are equal to 1 for the authentic images (corresponding to the images of a specific pose) and 0 for the imposter images (corresponding to the images of the other poses), expressed as

$$\mathbf{Y}^T \mathbf{F} = \mathbf{c} \quad (2)$$

where $\mathbf{Y}=[\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N]$ is the 1D Fourier transform of the low-dimensional feature obtained by PCA; $\mathbf{c}=[c_1, c_2, \dots, c_N]$ is a $N \cdot 1$ vector and c_t denotes the correlation peak amplitude of the t -th training image; c_t is equal to 1 for the authentic images and 0 for the imposter images.

Therefore, the objective of the MACE filter is:

$$\min_{\mathbf{F}} \mathbf{F}^T \mathbf{Q} \mathbf{F}, \text{ s.t. } \mathbf{Y}^T \mathbf{F} = \mathbf{c} \quad (3)$$

As a result, the solution of the above objective function is obtained by using the method of Lagrange multipliers. The optimum solution of MACE is

$$\mathbf{F}_{MACE} = \mathbf{Q}^{-1} \mathbf{Y} (\mathbf{Y}^T \mathbf{Q}^{-1} \mathbf{Y})^{-1} \mathbf{c} \quad (4)$$

The solution of the 1D-MVSDF filter is derived in the same way as the 1D-MACE filter. The optimum solution is

$$\mathbf{F}_{MVSDF} = \mathbf{G}^{-1} \mathbf{Y} (\mathbf{Y}^T \mathbf{G}^{-1} \mathbf{Y})^{-1} \mathbf{c} \quad (5)$$

where \mathbf{G} is an identity matrix if the input noise is modeled as white noise.

Based on the combination of (4) and (5), 1D-OTF is written as:

$$\mathbf{F}_{OTF} = \mathbf{T}^{-1} \mathbf{Y} (\mathbf{Y}^+ \mathbf{T}^{-1} \mathbf{Y})^{-1} \mathbf{c} \quad (6)$$

where $\mathbf{T} = \alpha \mathbf{Q} + (1 - \alpha) \mathbf{G}$, where α ($0 \leq \alpha \leq 1$) is a trade-off parameter: $\alpha=0$ leads to the 1D-MVSDF filter and $\alpha=1$ leads to the 1D-MACE filter.

3 Experimental Results and Analysis

In this section, three popular face databases with large pose variations, including PIE [20], HPI [21], and UMIST [22], are used to demonstrate the effectiveness of the proposed method.

The PIE face database contains 41,368 images of 68 different persons with variations in pose, illumination, and expression. We choose 612 images with three different yaw poses, that is, left-profile ($[-90^\circ, -15^\circ]$), frontal ($[-15^\circ, 15^\circ]$), and right-profile ($[15^\circ, 90^\circ]$). And each pose category has nine images for every person. The HPI face database has 15 persons with various poses. We choose ten images for both the left-profile and right-profile poses and six images for the frontal pose for each person. The UMIST face database consists of 564 different poses images of 19 persons. We respectively choose six images for each pose in our experiments.

All the faces in the images are cropped and resized to the size of 64×64 . For all the databases, we randomly choose 30% of the images as the training set and the rest are used as the test set. The experiments are repeated 30 times. We report the average pose/face classification rates obtained by the competing methods.

We show the head pose estimation results obtained by the other competing methods (include PCA [23], LDA [24], QRLDA [25] and the OTF method [26]) and our proposed DCFS method. To show the effectiveness of the proposed method, we use three different face representations, that is, gray, Gabor [27], and HOG [28] features. Table 1 shows the results of the competing methods for head pose estimation based on the three features.

From Table 1, we can see that the proposed method obtains the best performance for the three different face representations for most cases, which shows the feasibility and robustness of DCFS. In particular, DCFS achieves 100% accuracy on the PIE and UMIST databases. Most methods using the HOG feature can achieve higher pose estimation performance than those using the gray and Gabor features. This is because that HOG uses the histogram of gradients to describe the shape of a face image, which makes HOG contain the information of face direction. Therefore, the HOG feature is more effective for head pose estimation. In contrast, the performance of the Gabor feature is worse than HOG because Gabor extracts features that are insensitive to the variations caused by pose.

To demonstrate the superiority of the proposed DCFS for head pose estimation, we further show the distance distributions between the test images and the templates for all the competing methods on the PIE database (see Fig. 2).

Table 1. The results obtained by different competing methods for head pose estimation feature based on three representations

Method	Feature	Face Database		
		PIE	HPI	UMIST
PCA	Gabor	43.73%	75.71%	32.74%
	Gray	57.56%	78.04%	52.34%
	HOG	89.33%	89.68%	86.55%
LDA	Gabor	62.20%	75.60%	61.02%
	Gray	62.20%	82.34%	64.09%
	HOG	93.54%	90.64%	87.28%
QRLDA	Gabor	60.62%	70.75%	16.04%
	Gray	77.05%	76.00%	48.48%
	HOG	94.98%	78.00%	60.91%
CFA	Gabor	73.63%	56.09%	6.47%
	Gray	93.23%	64.36%	46.74%
	HOG	93.94%	87.18%	84.54%
DCFS	Gabor	75.08%	87.09%	79.37%
	Gray	94.11%	76.55%	70.56%
	HOG	100.00%	99.61%	100.00%

Figure 2 shows that PCA is not suitable for head pose estimation since the distances between different pose images are not big enough to be distinguished. Although PCA effectively reduces the noise, the extracted features are not distinguishable for pose estimation. LDA is better than PCA since LDA considers the class label information. LDA and QRLDA achieve a similar performance due to these methods try to differentiate all the classes. In contrast, OTF gives the better distance distributions than LDA and QRLDA. However, compared to OTF, DCFS shows the wonderful ability to separate the distance distributions for different templates, since a specific filter is designed to classify between one profile and the other two in the PCA feature subspace, which makes the head pose estimation more effective and robust.

4 Conclusions

In this paper, we propose a novel feature extraction framework called DCFS for robust head pose estimation. A novel principal optimal tradeoff filter, called POTF, is designed by using the frequency representations of 1D features on the feature subspace obtained by principal component analysis (PCA). Experimental results show that the DCFS with the HOG feature achieves great performance on head pose estimation. In our future work, we plan to extend the head pose estimation to handle more types of pose.

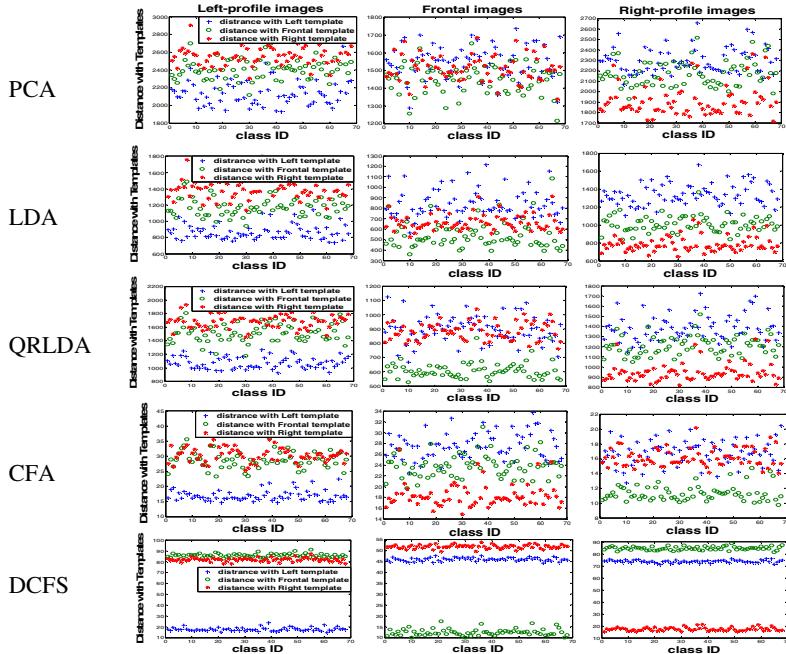


Fig. 2. Distance distributions between the test images and the templates. X axis represents different persons (i.e., class ID) while Y axis means the distance between a test image and the templates. The blue points ('+') indicate the distances between the test images and the left-profile template; green points ('o') show the distances between the test images and the frontal template; and red points ('*') represent the distances between the test images and the right-profile template. The test images are left-profile images (left column), frontal images (middle column), and right-profile images (right column).

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 61201359 and 61170179, by the Natural Science Foundation of Fujian Province of China under Grant 2012J05126, by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20110121110033.

References

1. Zhao, W.Y., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. *ACM Comput. Surv.* 35(4), 399–458 (2003)
2. Murphy-Chutorian, E., Trivedi, M.: Head Pose Estimation in Computer Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(4), 607–626 (2009)
3. Cootes, T.F., Walker, K., Taylor, C.J.: View-Based Active Appearance Model. In: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition, pp. 227–231 (2000)
4. Dantone, M., Gall, J., Fanelli, G., Gool, L.V.: Real-time Facial Feature Detection using Conditional Regression Forests. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2578–2585 (2012)

5. Ho, H.T., Chellappa, R.: Automatic Head Pose Estimation Using Randomly Projected Dense SIFT Descriptors. In: Proc. Int'l Conf. Image Processing, pp. 153–156 (2012)
6. Lowe, D.: Distinctive Image Features from Scale-Invariant Key points. *Int. J. Comput. Vision* 60(2), 91–110 (2004)
7. Vijaya Kumar, B.V.K., Savvides, M., Xie, C.: Correlation Pattern Recognition for Face Recognition. *Proc. IEEE* 94(11), 1963–1976 (2006)
8. Cootes, T., Wheeler, G., Walker, K., Taylor, C.: View-Based Active Appearance Models. *Image Vision Comput.* 20(9–10), 657–664 (2002)
9. Ji, Q., Hu, R.: 3D Face Pose Estimation and Tracking from A Monocular Camera. *Image Vision Comput.* 20(7), 499–511 (2002)
10. Srinivasan, S., Boyer, K.: Head Pose Estimation Using View Based Eigenspaces. In: Proc. Int'l Conf. Pattern Recognition, pp. 302–305 (2002)
11. Wei, Y., Fradet, L., Tan, T.: Head Pose Estimation Using Gabor Eigenspace Modeling. In: Proc. Int'l Conf. Image Processing, 281–284 (2002)
12. Kim, K.H., Zhang, C., Zhang, Z., Choi, S.: Robust Part-based Face Matching with Multiple Templates. In: Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (2013)
13. Milanfar, P., Cruz, S.: A Tour of Modern Image Filtering: New Insights and Methods, Both Practical and Theoretical. *IEEE Signal Processing Magazine* 1, 106–128 (2013)
14. Mahalanobis, A., Vijaya Kumar, B.V.K., Casasent, D.: Minimum Average Correlation Energy Filters. *Appl. Opt.* 26(17), 3630–3633 (1987)
15. Vijaya Kumar, B.V.K.: Minimum Variance Synthetic Discriminant Functions. *J. Opt. Soc. Amer. A* 3, 1579–1584 (1986)
16. Refregier, P.: Filter Design for Optical Pattern Recognition: Multi-Criteria Optimization Approach. *Opt. Lett.* 15(15), 854–856 (1990)
17. Venkataraman, K., Vijaya Kumar, B.V.K.: Performance of Composite Correlation Filters for Fingerprint Verification. *J. Opt. Engineering* 24(8), 1820–1827 (2004)
18. Hennings, P., Vijaya Kumar, B.V.K.: Palmprint Recognition Using Correlation Filter Classifiers. In: Proc. Signals, Systems and Computers, pp. 567–571 (2004)
19. Park, S., Smith, M.J.T., Mersereau, R.M.: A new directional filter bank for image analysis and classification. In: Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, pp. 1417–1420 (1999)
20. Sim, T., Baker, S., Bsat, M.: The CMU Pose Illumination and Expression (PIE) Database of Human Faces. In: Proc. AFGR, pp. 46–51 (2002)
21. Gourier, N., Hall, D., Crowley, J.L.: Estimating Face Orientation from Robust Detection of Salient Facial Features. In: Proc. ICPR Pointing (2004)
22. <http://www.sheffield.ac.uk/eee/research/iel/research/face>
23. Turk, M.: Eigenfaces for Recognition. *J. Cogn. Neurosci.* 3(1), 71–86 (1991)
24. Yang, J., Frangi, A.F., Yang, J.Y., Zhang, D.: KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(2), 230–244 (2005)
25. Ye, J., Li, Q.: A Two-Stage Linear Discriminant Analysis via QR-Decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(6), 929–941 (2005)
26. Yan, Y., Wang, H.Z., Li, C.H., Yang, C.H., Zhong, B.N.: A Novel Unconstrained Correlation Filter and Its Application in Face Recognition. In: Proc. Workshop on Intelligence Science and Intelligent Data Engineering, pp. 32–39 (2013)
27. Xie, S., Shan, S., Chen, X., Chen, J.: Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition Image Processing, vol. 19, pp. 1349–1361 (2010)
28. Navneet, D., Bill, T.: Histograms of Oriented Gradients for Human Detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)

Automated Tongue Segmentation Based on 2D Gabor Filters and Fast Marching

Zhenchao Cui¹, Wangmeng Zuo¹, Hongzhi Zhang¹, and David Zhang^{1, 2}

¹ School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

² Biometrics Research Centre, Department of Computing,

Hong Kong Polytechnic University, Hong Kong

{cui zhenchao, cswm zuo, zhanghz0451}@gmail.com,

csdzhang@comp.poly.edu.cn

Abstract. In this paper, we propose a novel automated tongue segmentation scheme which can well address the edge enhancement and the initialization problem of the tongue body contour. First, according to the grey level distribution of the contour, we propose a 2D Gabor magnitude - based detector for the enhancement of the contour of tongue body. Second, to cope with the discontinuity of edge detection results, we select two stable segments of the tongue body contours and use the fast marching method to obtain the closed tongue body contour. Moreover, gradient vector flow (GVF) snake is used to obtain the final segmentation result and an augmented Lagrangian method is adopted for fast computation of GVF field. Qualitative and quantitative comparisons further verify the superiority of the proposed method for the segmentation of tongue body.

Keywords: Segmentation, computerized tongue diagnosis, fast marching, 2D Gabor filter, active contour.

1 Introduction

Since of its convenient and non-invasive nature, computerized tongue diagnosis has been gradually received considerable research interests, and several systems have been reported. Using color and texture features extracted from tongue images, Pang et al. employed Bayesian network method to develop a computerized tongue inspection system [1]. Gao et al. [2] developed a support vector machine (SVM)-based computerized tongue diagnosis system.

Tongue body segmentation is a prerequisite to the computerized tongue diagnosis system. Generally, tongue body segmentation usually involves two major steps: edge enhancement and detection of the tongue body contour. For edge enhancement, several conventional edge detectors [3, 4] have been adopted for the enhancement of the tongue contours. Considering the shape and grey level characteristics of tongue contour, Zuo et al. [5, 6] proposed a polar edge detector to effectively suppress the adverse interference from the lip boundary, tongue fissures, etc. For the detection of the tongue body contour, deformable models [3], active contour models (i.e., snake)

[3, 5], and GVF snakes have been proposed for tongue body segmentation. For these models, initialization of the contour usually plays a critical role on the final segmentation results. Pang et al. [3] suggested a bi-elliptical deformable template to utilize both the edge enhancement result and the shape of tongue body for the automated initialization of the contour.

Although the progress has been made in edge enhancement and tongue body contour detection, automated segmentation of tongue body remains a challenging problem. First, for edge enhancement, the existing methods are based on conventional edge detectors, and neglect the characteristics of gray level variation of the real contour of tongue body. Second, the detection of the tongue body contour also suffers from the weak edges of tongues and the interference from lip and tongue fissures.

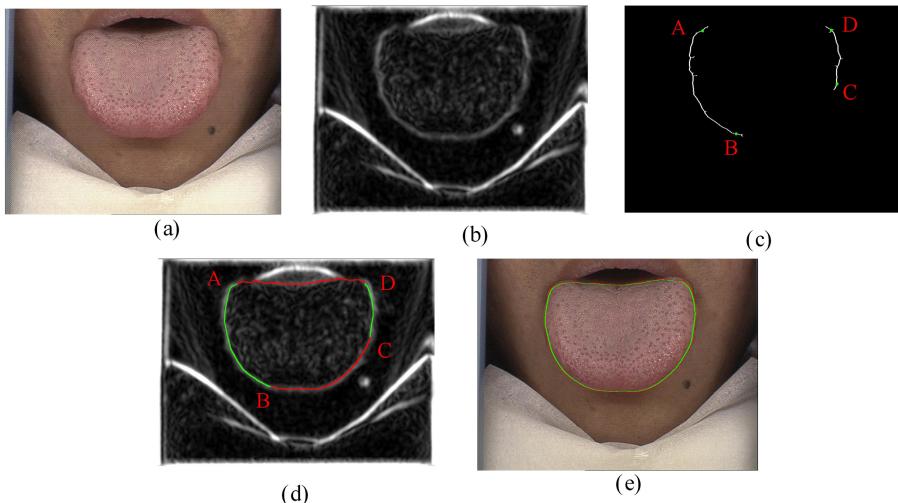


Fig. 1. Results of the proposed segmentation method: (a) a tongue image, (b) results after edge enhancement, (c) result after edge thresholding and the selection of the two stable segments, where the points A, C, B, D are four points in the two segments; (d) the result of fast marching, and (e) the segmentation results, where the red curve is the initial contour, and the green curve the final contour.

In this paper, considering the characteristics of the tongue body contour and the weak edge of tongue, we propose a novel tongue body segmentation method by using 2D Gabor filters, fast marching, and GVF snake. As shown in Fig. 1, the proposed method first uses the 2D Gabor filters for edge enhancement. After edge thresholding, to alleviate the edge discontinuity problems, we select two stable segments, and use the fast marching method to obtain a continuous contour for initialization. Finally, the active contour model is adopted for tongue body segmentation. For simplicity, hereafter we call the proposed method as the GaborFM method.

The remainder of the paper is organized as follows. Section 2 gives the 2D Gabor magnitude - based edge detector. Section 3 describes the proposed scheme for the segmentation of tongue body. Section 4 provides the qualitative and quantitative results. Finally, Section 5 offers our conclusion.

2 2D Gabor Magnitude – Based Edge Detection

2.1 2D Gabor Magnitude - Based Detector

Fig. 2 shows the typical profiles of the boundary pixel of the tongue body. For the typical four boundary pixels shown in Fig. 2(a), Fig. 2(b) shows the profiles of the intensities along the white lines, and Fig. 2(c) shows the profiles of the real and the imaginary parts of the Gabor filter. From Fig. 2, the profiles of the boundary pixel are similar with either the real or the imaginary parts of the Gabor filter. So it is proper to use 2D Gabor filters for the enhancement of the boundary of tongue body.

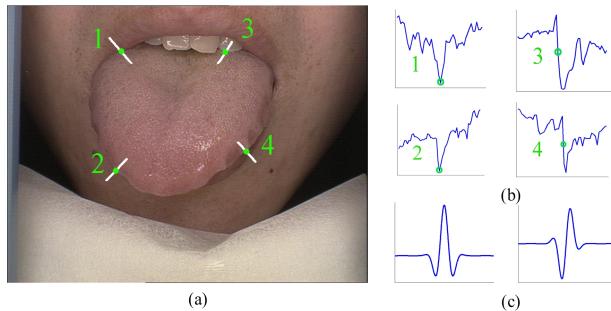


Fig. 2. Typical profiles of the boundary pixel of the tongue body: (a) Tongue image, (b) profiles of the intensities of the four boundary pixels along the white lines in (a), and (c) profiles of the real and the imaginary parts of the Gabor filter.

Here we adopt the 2D Gabor function derived by Lee [7],

$$\psi(x, y, x_0, y_0, \omega, \theta, \kappa) = \frac{\omega}{\sqrt{2\pi}\kappa} e^{-\frac{\omega^2}{8\kappa^2}(4x'^2+y'^2)} \left(e^{j\omega x'} - e^{-\frac{\kappa^2}{2}} \right) \quad (1)$$

where $x' = (x - x_0)\cos\theta + (y - y_0)\sin\theta$, $y' = (x - x_0)\sin\theta + (y - y_0)\cos\theta$, (x_0, y_0) is the center of the function, ω is the radial frequency in radians per unit length, and θ is the orientation of the Gabor functions in radians. The κ is defined by $\kappa = \sqrt{2\ln 2}(2^\delta + 1)/(2^\delta - 1)$, where δ is the half-amplitude bandwidth of the frequency response, which is between 1 and 1.5 octaves according to neurophysiological findings [11]. When ω and δ are fixed, σ can be derived from $\sigma = \kappa/\omega$.

We propose a 2D Gabor magnitude - based detector for edge enhancement. We choose $\omega = 0.62$, $\delta = 1$, and use $G_k(x, y)$ to denote the Gabor filter with the orientation of $\theta = k/8\pi$ ($k = 0, 1, \dots, 7$). Given a tongue image $I(x, y)$, the convolution of the image I and G_k is,

$$FI_k = I * G_k \quad (2)$$

where $*$ denotes the convolution operator. 2D Gabor magnitude-based detector is defined as:

$$M_{\max}(x, y) = \max_k \sqrt{FI_k(x, y) \cdot FI_k(x, y)} \quad (3)$$

Fig. 3 shows an example of the edge enhancement results on one typical tongue image. Compared with the Sobel, Canny, and derivative of Gaussian (DoG) methods, one can see that our 2D Gabor magnitude – based detector is more effective for the edge enhancement of tongue image.

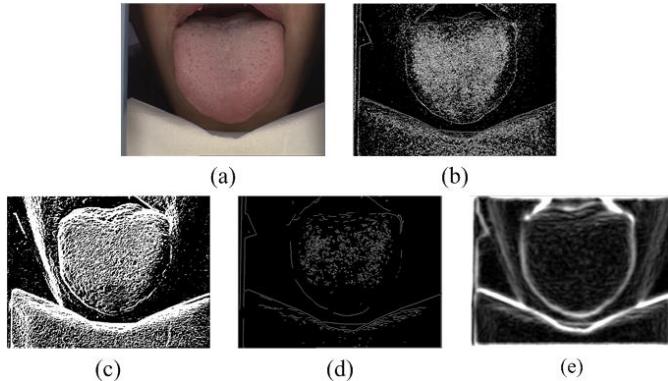


Fig. 3. Enhancement of the boundary of tongue body using different methods: (a) original image, (b) Sobel operator, (c) DoG filter, (d) Canny operator, and (e) 2D Gabor magnitude

2.2 Edge Thresholding

Generally, one typical tongue image usually involves three major components: background, tongue body, and other facial parts, as shown in Fig. 4 (b). Since of the stable and uniform lighting condition, the background would be stable and can be easily segmented from the tongue image. Here we first transform the color tongue image into the YIQ color space, and use the Otsu's method [8] to determine a threshold T_b for the I-channel of the YIQ color space. The reason to choose I channel of YIQ is that increasing I characterizes the change from blue, through purple, to red colors, and I is more effective to separate background pixels from the others than Y and Q. Then we set the pixels with the I value smaller than T_b as background pixels, and use morphological operators, i.e., dilation, filling, and erosion, to refine it.

We further use Q-channel to mask the parts of non-boundary pixels. Let T_l denote the threshold obtained by the Otsu's method. We choose the threshold $T_Q = 2.1T_l$ and set the pixels with the Q value higher than T_Q as tongue body pixels. Similarly, morphological operators are then used to refine the tongue body region. Actually, the method above can obtain a subset of the background and tongue body components, we directly assign these pixels as non-boundary pixels, and modify $M_{\max}(x, y) = 0$ for these pixels, as shown in Fig. 4(c).

Finally, we define a threshold T_M for the binarization of the edge image. Let T_O be the threshold obtained by the Otsu's method on $M_{max}(x, y)$, and Var_M be the standard variance of $M_{max}(x, y)$. The threshold T_M is then defined as:

$$T_M = \begin{cases} T_o, & \text{if } T_o > 1.1Var_M \\ 1.1Var_M, & \text{else} \end{cases} \quad (4)$$

After edge thresholding, we employ the morphological operators to obtain single-pixel edge curves, resulting in the final binarized edge image, as shown in Fig. 4(d).

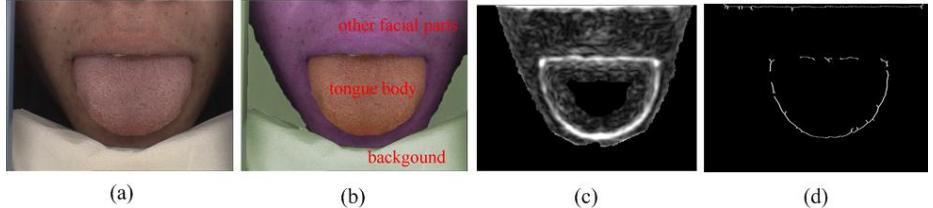


Fig. 4. Thresholding of the edge image: (a) original image, (b) three major components of a tongue image, (c) edge image after masking out parts of the non-boundary pixels, and (d) the final binarized edge image.

3 Contour Detection Using Fast Marching and Active Contour Model

3.1 Selection of Stable Segments

Based on the estimated center of the tongue body (\bar{x}, \bar{y}) , we divide the binarized edge image into the left and the right parts. For each part, we extract the segment with the largest length as a stable segment. Moreover, to verify the stable segments, we further consider the approximate symmetry of the two segments. Fig. 5(b) shows an example of the two stable segments. Finally, for each stable segment, we back-track the two end points to find two stable points, as shown in Fig. 5(b).

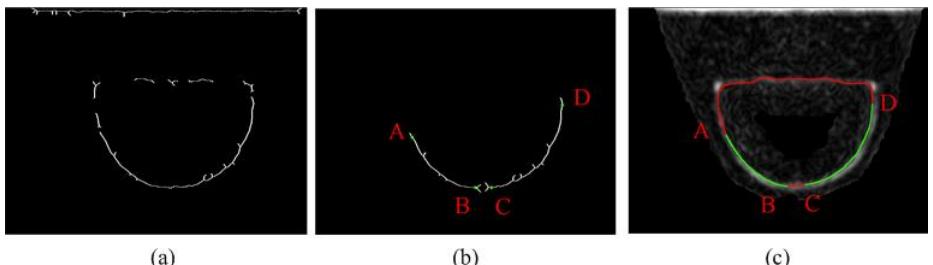


Fig. 5. Selection of stable segments: (a) the binarized edge image, (b) the stable segments with the corresponding stable points, and (c) the result of fast marching

3.2 Contour Initialization Using Fast Marching

For the initialization of tongue body contour, we utilize the fast marching method to address the discontinuity problem. Given two points $A(x_A, y_A)$ and $C(x_C, y_C)$, we define the length of a planar curve $\gamma: [0, 1] \mapsto \mathbb{R}^2$ from A to C,

$$\mathcal{L}_F(\gamma) = \int_0^1 g(\gamma(t)) \|\gamma'(t)\|_2 dt \quad (5)$$

where $\|\gamma'(t)\|_2$ denotes the l_2 -norm of the gradient of $\gamma(t)$, $g(x, y)$ is defined in \mathbb{R}^2 .

$$g(x, y) = \frac{1}{1 + M_{\max}(x, y)} \quad (6)$$

Then we defined the required curve from A to C is a shortest path defined by $\mathcal{L}_F(\gamma)$,

$$\gamma^*(t) = \arg \min_{\gamma(t) \in \mathcal{C}} \mathcal{L}_F(\gamma(t)) \quad (7)$$

where \mathcal{C} is the set of curves with $\gamma(0) = (x_A, y_A)$ and $\gamma(1) = (x_C, y_C)$.

According to [9, 10], γ^* can also be estimated by solving the following Eikonal equation,

$$\|\nabla T_A(x, y)\| = g(x, y) \quad (8)$$

with $T_A(x_A, y_A) = 0$, where $T_A(x, y)$ denotes the shortest distance \mathcal{L}_F of (x_A, y_A) and (x, y) . Here the fast marching method is adopted to solve the Eikonal equation. To obtain $T_A(x, y)$ efficiently, we stop fast marching when the point (x_C, y_C) is reached.

To avoid the interference of lip, we find two paths for any two points. If the difference in length \mathcal{L}_F of the two paths is low, we choose the inner curve; else we choose the one with the lower length. Fig. 1(d) and Fig. 5(c) show two examples of the shortest paths constructed by fast marching method, and satisfactory initialization results can be obtained.

3.3 Gradient Vector Flow Snake

We adopt the gradient vector flow (GVF) snake model [11] for the final tongue body segmentation. Given an edge image $M(x, y)$, the GVF field $\mathbf{w}(x, y) = [u(x, y), v(x, y)]$ can be obtained by solving the minimization problem,

$$E(\mathbf{w}(x, y)) = \iint \infty |\nabla \mathbf{w}|^2 + |\nabla M|^2 |\mathbf{w} - \nabla M|^2 dx dy \quad (9)$$

where $|\cdot|$ denotes the l_2 norm with $|\nabla \mathbf{w}|^2 = |u_x^2 + u_y^2 + v_x^2 + v_y^2|$. For fast GVF computation, we use the augmented Lagrangian method (ALM) - based algorithm proposed in [12].

For GVF snake, we have $\nabla E_{ext} = \mathbf{w}(x, y)$. So, after the computation of the GVF field, we can evolve the curve dynamically until convergence, and the partial derivative of the curve $\mathbf{x}(s, t)$ with respective to time t as,

$$\mathbf{x}'_t(s, t) = \alpha \mathbf{x}''(s, t) - \beta \mathbf{x}'''(s, t) - \mathbf{w} \quad (10)$$

For tongue body segmentation, we choose $\alpha = 0.1$, $\alpha = 0.05$, and $\beta = 0$.

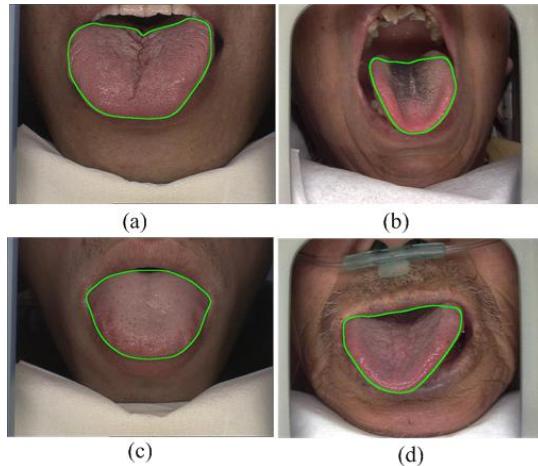


Fig. 6. The segmentation results

4 Experimental Results

We first use four typical tongue images with various shapes, texture, color and interference to evaluate GaborFM. All the images are acquired by our tongue image acquisition device with semi-enclosed environment and stable lighting condition. The image size is 768×576. Fig. 6 shows the segmentation results obtained using the proposed method, and GaborFM can correctly segment these images.

Then, we use a tongue image data set of 300 images to quantitatively evaluate the proposed method. We manually segment tongue body from each image, and use the segmentation results as the ground truth, and adopt boundary- and area-based criteria for evaluation. In boundary-based evaluation [13], the Hausdorff distance (HD) and the mean distance to the closest point (MD) are used. In area-based evaluation [14], the false negative volume fraction (FNVF, %) and the false positive volume fraction (FPVF, %) are adopted.

Table 1. The HD, MD distances, FPVF and FNVF of BEDC, PolarSnake and GaborFM

Method	HD	MD	FPVF(%)	FNVF(%)
BEDC [3]	49.59±36.91	21.80±10.66	13.89±5.63	19.89±15.51
PolarSnake [5]	30.35±22.07	8.57±4.34	1.93±3.94	10.04±11.56
GaborFM	21.59±13.25	6.85±2.96	1.38±3.42	8.47±11.00

Table 1 lists the quantitative evaluation results. GaborFM can obtain lower HD and MD distance than BEDC and PolarSnake and the standard deviation (std.) of HD and MD distance of GaborFM are lower than BEDC and PolarSnake, which indicates that the proposed method is better than BEDC and PolarSnake in terms of boundary-based performance criteria. GaborFM can also achieve lower FPVF and FNVF than BEDC and PolarSnake, which indicates that the proposed method is superior to BEDC and PolarSnake in terms of area-based performance criteria.

5 Conclusion

We propose a GaborFM method for tongue body segmentation. Considering the characteristics of the tongue body contour, a Gabor magnitude-based detector is developed for edge enhancement. We further take into account both the color characteristics and the edge enhancement result for edge image thresholding. Then we select two stable segments, and use fast marching for initialization. Finally, GVF snake is used for tongue body segmentation and ALM is adopted for fast GVF computation. The proposed method can well address the edge enhancement and the contour discontinuity problems. Experimental results show that the proposed method is superior to the existing tongue body segmentation methods, i.e., BEDC [3] and PolarSnake [5, 6].

References

1. Pang, B., Zhang, D., Li, N.M., Wang, K.Q.: Computerized Tongue Diagnosis Based on Bayesian Networks. *IEEE Trans. Biomedical Engineering* 51, 1803–1810 (2004)
2. Gao, Z., Cui, M., Lu, G.M.: A Novel Computerized System for Tongue Diagnosis. In: International Seminar on FITME, pp. 364–367 (2008)
3. Pang, B., Zhang, D., Wang, K.Q.: The Bi-Elliptical Deformable Contour and Its Application to Automated Tongue Segmentation in Chinese Medicine. *IEEE Trans. Medical Imaging* 24, 946–956 (2005)
4. Yu, S.Y., Yang, J., Wang, Y.G., Zhang, Y.: Color Active Contour Models Based Tongue Segmentation in Traditional Chinese Medicine. In: 1st ICBBE, pp. 1065–1068 (2007)
5. Zuo, W.M., Wang, K.Q., Zhang, D., Zhang, H.Z.: Combination of Polar Edge Detection and Active Contour Model for Automated Tongue Segmentation. In: 3rd International Conference on Image and Graphics, pp. 270–273 (2004)
6. Zhang, H.Z., Zuo, W.M., Wang, K.Q., Zhang, D.: A Snake-Based Approach to Automated Segmentation of Tongue Image Using Polar Edge Detector. *International Journal of Imaging Systems and Technology* 16(4), 103–112 (2006)
7. Lee, T.S.: Image representation using 2D Gabor wavelet. *IEEE Trans. Pattern Analysis and Machine Intelligence* 18(10), 959–971 (1996)
8. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. System, Man, and Cybernetics* 9(1), 62–66 (1979)
9. Sethian, J.A.: *Level Set Methods and Fast Marching Methods*. Cambridge University Press (1999)
10. Pechaud, M., Keriven, R., Peyre, G.: Extraction of Tubular Structures over an Orientation Domain. In: CVPR, Miami, pp. 336–342 (2009)
11. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Processing* 7(3), 359–369 (1998)
12. Li, J.F., Zuo, W.M., Zhao, X.F., Zhang, D.: An augmented Lagrangian method for fast gradient vector flow computation. In: ICIP 2011, pp. 1557–1560 (2011)
13. Chalana, V., Kim, Y.: A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Trans. Medical Imaging* 16(5), 642–652 (1997)
14. Udupa, J.K., LeBlanc, V.R., Schmidt, H., Imielinska, C., et al.: A methodology for evaluating image segmentation algorithms. In: SPIE vol. 4684, pp. 266–277 (2002)

Hyperspectral Medical Images Unmixing for Cancer Screening Based on Rotational Independent Component Analysis

Bo Du¹, Nan Wang², Liangpei Zhang², and Dacheng Tao³

¹ School of Computer, Wuhan University

² LIESMARS, Wuhan University

³ Centre for Quantum Computation and Intelligent Systems

University of Technology , Sydney

gunspace@163.com

Abstract. Hyperspectral images have shown promising performance in many applications, especially extracting information from remotely sensed geometric images. One obvious advantage is its good ability to reflect the physical meaning from a point view of spectrum, since even two very similar materials would present an obvious difference by a hyperspectral imaging system. Recent work has made great progress on the hyperspectral fluorescence imaging techniques, which makes the elaborate spectral observation of cancer areas possible. Cancer cells would be distinguishable with normal ones when the living body is injected with fluorescence, which helps organs inside the living body emit lights, and then the signals can be obtained by the passive imaging sensor. This paper discusses the ability to screen the cancers by means of hyperspectral bioluminescence images. A rotational independent component analysis method is proposed to solve the problem. Experiments evaluate the superior performance of the proposed ICA-based method to other blind source separation methods: 1) The ICA-based methods do perform well in detect the cancer areas inside the living body; 2) The proposed method presents more accurate cancer areas than other state-of-the-art algorithms.

Keywords: Cancer detection, hyperspectral images, independent component analysis.

1 Introduction

Much efforts have been done to combine the advantages of bioluminescence and fluorescence imaging [1, 2]. Actually, multispectral *in vivo* optical imaging technologies with bioluminescence and fluorescence is drawing great interest in recent years [3, 4, 5]. It employs bioluminescence and fluorescence imaging to obtain useful signals of receptors and has become another important biomedicine imaging techniques. Compared with the conventional biomedicine imaging techniques, such as ultrasonic, Computed tomography, Magnetic Resonance Imaging, positron emission tomography, it provides more straightforward measurements, much safer performance

and lower costs. One significant progress is the Maestro imaging system [6-8], which can obtain the hyperspectral images, covering the visible to near infrared with fine spectral resolution.

As to the hyperspectral image, it is a powerful way to describe the physical meaning of different materials [9]. Its foremost advantage is that the spectral resolution is very fine and the corresponding spectrum of each different material is continuous and smooth, showing diagonal features for elaborately separating visually very similar objects. So hyperspectral images provides a new way to analyze the distribution of materials of interest [10], which have been widely used in geometrical information extraction. Classical methods employ the spectral unmixing to do the task. It is assumed that the pixels in the image are linear composed of limited materials' spectra (called endmembers) and the corresponding abundances. This is the so called linear mixture model (LMM) [11, 12, 13]. There are two ways for spectral unmixing: getting the typical spectra for each material and then the abundances can be obtained by least squares methods with these spectra; unmixing the pixels into the spectra and the abundances simultaneously [14, 15, 16]. The latter approach is usually achieved by blind source separation based methods, including independent component analysis and nonnegative matrix fraction methods [17-23]. But, these ICA-based methods cannot obtain nonnegative abundances which is necessary in explaining the objects' distributions and NMF are susceptible to the initial values.

This paper aims to address the cancer screening problem from a hyperspectral fluorescence images. A rotational independent component analysis is developed to hyperspectral unmixing. The rotational ICA rotates the coordinate system of the dataset with a serial of orthogonal rotation matrix until all the data fall in the first quadrant. Compared with traditional ICA, the result of the rotational ICA can make most abundances values non-negative, satisfying the abundance non-negative constraint, which is an important property in reality. With the accurate abundance vectors, reliable endmembers can also be obtained. Besides, the proposed rotational ICA can be achieved without a proper initialization.

The remainder of this paper is organized as follows. Section 2 presents the LMM and basic ICA model. Section 3 details the rotational ICA method and applies it to hyperspectral unmixing. The experiments on the fluorescence dataset are described in Sections 4, respectively. Section V concludes the paper.

2 Spectral Unmixing Model in Hyperspectral Images

2.1 A Linear Mixture Model (LMM)

The LMM assumes that one pixel in the hyperspectral dataset is a linear mixture of P known material signatures, called endmembers: $\mathbf{A} = [a_1, a_2, \dots, a_p]$, where a_i is one of the endmember spectra with dimension "band". The corresponding proportion is called the abundance: $\mathbf{S} = [s_1^T, s_2^T, \dots, s_p^T] = [\omega_1, \omega_2, \dots, \omega_N]$, where each column s_i^T is a N -dimension vector, corresponding to the i th spectra in A . Based on LMM, each pixel in a hyperspectral image dataset can be expressed as :

$$x = A\omega + \varepsilon \quad (1)$$

where x is a $band \cdot 1$ vector representing one pixel in the hyperspectral image, and ε is the residual error. According to LMM, the abundance matrix should satisfy the ASC and abundance non-negative constraints (ANC) simultaneously, i.e., $s_1^T + s_2^T + \dots + s_p^T = \mathbf{1}^T$ and $s_i^T \geq 0$

3 Rotational ICA for Hyperspectral Bioluminescence Images

In the traditional ICA, whitening is an important step before ICA iteration procedure [11]. Because it can achieve the half-work of the ICA, removing any second-order dependencies in the dataset. The whitened dataset always exists negative values (Figure 1). In order to make the whitened dataset fall in the first quadrant, which leads the results non-negative, the rotation is needed.

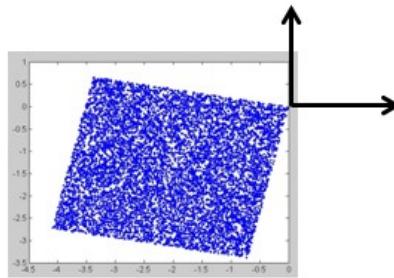


Fig. 1. The whitened dataset in two dimensions

The rotation matrix can be defined as:

$$W = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} \quad (2)$$

Through the rotation matrix, the coordinate system will be anticlockwise rotated and then some points will fall in the first quadrant. The matrix W is determined by the rotation angle ϕ and with the formula (4) W is constrained to be orthogonal. The new coordinate values with the rotation matrix W can be obtained as:

$$\begin{cases} y_1 = x_1 \cos \phi + x_2 \sin \phi \\ y_2 = x_2 \cos \phi - x_1 \sin \phi \end{cases} \quad (5)$$

To find the rotation matrix W , an objective function in 2-dimensions case is defined as:

$$\min J = \begin{cases} 0 & \text{if } y_1 \geq 0 \text{ and } y_2 \geq 0 \\ y_2^2 & \text{if } y_1 \geq 0 \text{ and } y_2 < 0 \\ y_1^2 & \text{if } y_1 < 0 \text{ and } y_2 \geq 0 \\ y_1^2 + y_2^2 & \text{otherwise} \end{cases} \quad (6)$$

where the function value is equal to zero when the point falls in the first quadrant, otherwise, equal to non-zero. Differentiating (6) with respect to the rotation angle ϕ , we get:

$$-\frac{\partial J}{\partial \phi} = \begin{cases} 0 & \text{if } y_1 \geq 0 \text{ and } y_2 \geq 0 \\ y_2 y_1 & \text{if } y_1 \geq 0 \text{ and } y_2 < 0 \\ -y_1 y_2 & \text{if } y_1 < 0 \text{ and } y_2 \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

We can minimize the problem (6) by finding a zero of the equation (7) with respect to the rotation angle ϕ . There are many standard methods available to find the zero of a function, a *fzero* in the Matlab function can be used to find the zero of the equation (7).

For n-dimension case, a general n-dimensional orthogonal transform can be formed from a product of 2-D rotations. In the iteration procedure, calculate the values of equation (7) for each axis pair and rotate the axis pair with the highest value.

Based on LMM, hyperspectral unmixing can be considered as a problem that extracts potential components from the observations. Consider the abundance vectors as independent components, we can perform hyperspectral unmixing based on rotational ICA. Because the abundance vectors are correlative each other, we whiten the original dataset with correlation matrix, not the covariance matrix, to keep the dataset correlative. The whole algorithm is as follows:

1) Whiten the original dataset X . Calculate the correlation matrix of original dataset:

$$\Sigma = XX^T / N \quad (8)$$

where N is the number of pixels. Calculate the eigenvalues matrix D and eigenvectors matrix E of the Σ , and reduce the dimension of original dataset to p -dimensions with

$$Z = (D_p^{-1/2} E_p^T) X \quad (9)$$

where D_p is a diagonal matrix with the first p eigenvalues in the diagonal line E_p is the corresponding vectors. Set $Z(0) = Z$ and $W(0) = I_n$ for $t = 0$.

- 2) Calculate the output abundances $Y = Z(t) = W(t)Z(0)$ and set Y_+ with $y_{ik}^+ = \max(y_{ik}, 0)$ and Y_- with $y_{ik}^- = \min(y_{ik}, 0)$
- 3) Calculate the values of equation (7) for all axis pairs

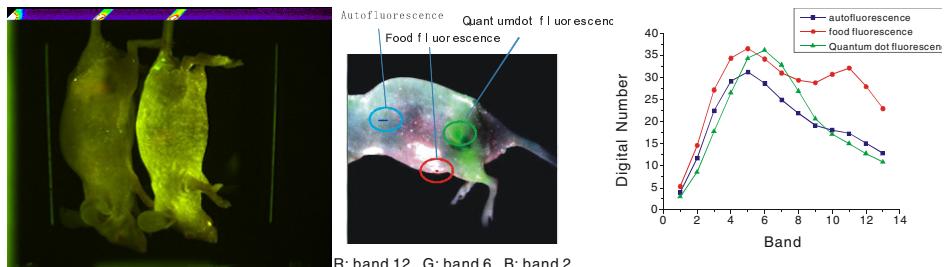
$$g_{ij} = \sum_k y_{ik}^+ y_{jk}^- - y_{ik}^- y_{jk}^+$$
- 4) If the maximum value of $|g_{ij}|$ is less than the tolerance required, stop. Otherwise, continue the following step.
- 5) Choose the axis pair i^*, j^* with max value in $|g_{ij}|$ and select rows i^*, j^* from matrix $Z(t)$ to construct the $2 \cdot p$ matrix Z^*
- 6) Using the reduced data Z^* to minimize the problem (6). Search a rotational angle ϕ to make the value of formula (7) be zero. This rotation angle is the stationary point of the function (6) which lead the extrema of the function (6).
- 7) Calculate the rotation matrix $R(t+1) = [r(t+1)_{ij}]$ with $\phi^*(t+1)$, where $r_{i^*i^*} = r_{j^*j^*} = \cos(\phi^*)$, $r_{j^*i^*} = -\sin(\phi^*)$, $r_{i^*j^*} = \sin(\phi^*)$, $r_{ii} = 1$ for all $i \neq i^*, j^*$ and all other entries of R are zero.
- 8) Set $W(t+1) = R(t+1)W(t)$ and $Z(t+1) = R(t+1)Z(t)$
- 9) Set $t = t + 1$ and go to step (2) until the max value of $|g_{ij}|$ is close to zero.
- 10) Repeating the step (2)-(9) in the procedure of non-negative ICA and get the abundance matrix S . According to X and S in the linear mixture model, calculate the endmember matrix A with the least squares estimation.

4 Experiments

In the experiment, three methods — FAST_ICA[11]、SISAL[9] and CICA[23] are employed to evaluate the efficient of the proposed strategy. The simplex identification via split augmented Lagrangian (SISAL) algorithm, which enforced the endmembers' spectral vectors to compose a convex hull containing all the pixels in the image, constrained by soft constraints. We use SISAL to extract the endmember signatures and calculate the abundance with least square algorithm.

The spectral angle distance (SAD) [2] is used to evaluate the accuracy of the extracted endmember signatures. The SAD values of the algorithms are showed in

Table 1 and the abundance maps are plotted in Figure 2. It is revealed that our proposed method presents the best endmembers, with the least SAD values for cancer areas and the other areas. From the abundance maps, it is also obvious that the cancer distribution is more accurate in our method.



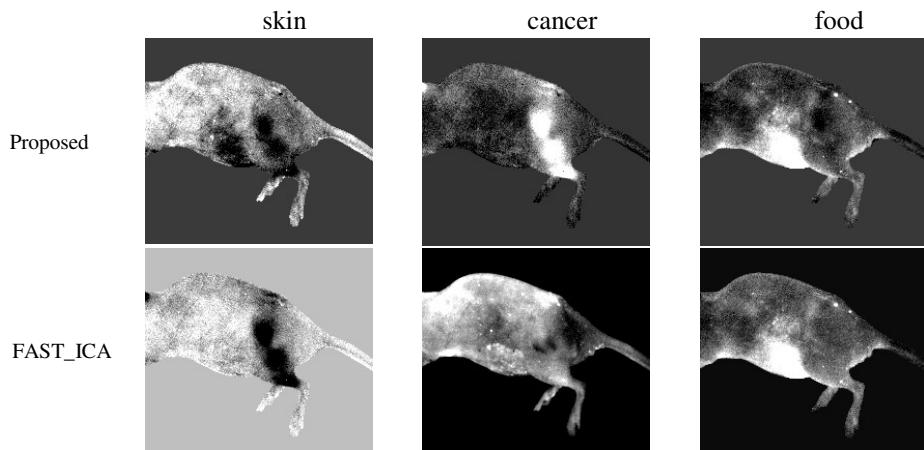
Hyperspectral Bioluminescence Datacube

The false color picture of the datacube in our experiments and the most concentrated areas for each fluorescence material

The reference spectra of three materials

Table 1. SAD values for different algorithm

	skin	cancer	food	average
Proposed	0.2048	0.0722	0.0282	0.1018
FAST_ICA	1.1062	0.0634	0.2342	0.4679
SISAL	2.8786	0.1211	0.0832	1.0277
CICA	0.2051	0.0815	0.0258	0.1041

**Fig. 2.** The abundance maps of different algorithms

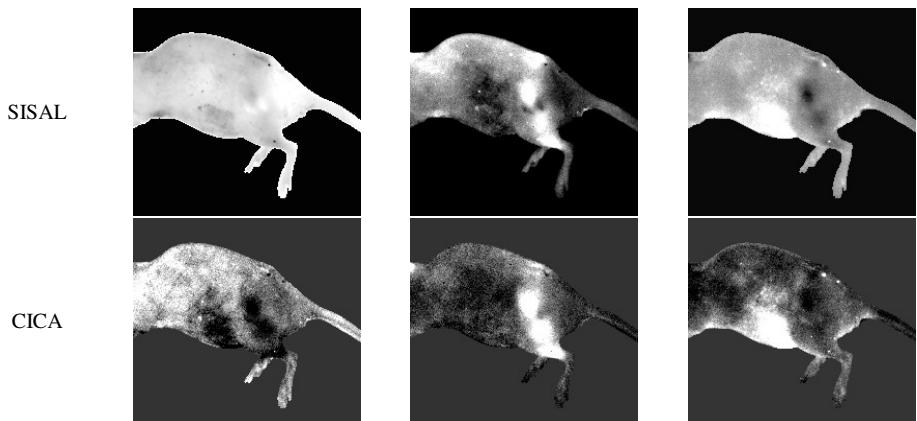


Fig. 2. (Continued.)

5 Conclusion

This paper proposes a rational based ICA, focusing on solving the cancer screening problem from hyperspectral medical images. By a rational transformation, most of the abundances values can be nonnegative and the corresponding spectra are also accurate . Experimental results show its superior performance than conventional methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 61102128, and 41061130553, the National Basic Research Program of China (973 Program) under Grant 2012CB719905 and 2011CB707105,

References

- [1] Wang, X., Rosol, M., Ge, S., et al.: Dynamic tracking of human hematopoietic stem cell engraftment using *in vivo* bioluminescence imaging. *Blood* 102, 3478–3482 (2003)
- [2] Levenson, R.M., Lynch, D.T., Kobayashi, H., et al.: Multiplexing with Multispectral Imaging: From Mice to Microscopy. *ILAR Journal* 49(1), 78–88 (2008)
- [3] Zhang, Y., Han, Y., Zhao, C.L., et al.: Current developments in animal *in vivo* optical imaging technologies with bioluminescence and fluorescence. *Chinese Bull. Life Sci.* 18(1), 25–30 (2006)
- [4] Mansfield, J.R., Hoyt, C., Levenson, R.M.: Visualization of Microscopy-Based Spectral Imaging Data from Multi-Label Tissue Sections. *Current Protocols in Molecular Biology* 14(19), 1–14 (2008)
- [5] Ntziachristos, V., Ripoll, J., Wang, L.V., et al.: Looking and listening to light the evolution of whole - body photonic imaging. *Nat. Biotechnol.* 23(3), 313–320 (2005)
- [6] Harris, S., Wallace, R.: Acousto-optic tunable filter. *Journal of the Optical Society of America* 59, 744–747 (1969)

- [7] Gebhart, S.C., Thompson, R.C., Mahadevan-Jansen, A.: Liquid-crystal tunable filter spectral imaging for brain tumor demarcation. *Applied Optics* 46, 1896–1910 (2007)
- [8] Cai, W., Chen, X.: Preparation of peptide-conjugated quantum dots for tumor vasculature-targeted imaging. *Nature Protocols* 3, 89–96 (2008)
- [9] Plaza, A., Martinez, P., Perez, R., Plaza, J.: Spatial/spectral endmember extraction by multidimensional morphological operations. *IEEE Trans. Geosci. Remote Sens.* 40(9), 2025–2041 (2002)
- [10] Bioucas-Dias, J.M.: A variable splitting augmented Lagrangian approach to linear spectral unmixing. In: First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS 2009, pp. 1–4 (2009)
- [11] Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*. Wiley, New York (2001)
- [12] Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* 13, 556–562 (2000)
- [13] Donoho, D., Stodden, V.: When Does Non-Negative Matrix Factorization Give a Correct Decomposition Into Parts (2003)
- [14] Zhou, G.X., Xie, S.L., Ding, S.X., Yang, J.M., Zhang, J.: Blind Spectral Unmixing Based on Sparse Nonnegative Matrix Factorization. *IEEE Trans. Image Process.* 20, 1112–1125 (2011)
- [15] Huck, A., Guillaume, M., Blanc-Talon, J.: Minimum Dispersion Constrained Nonnegative Matrix Factorization to Unmix Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* 48, 2590–2602 (2010)
- [16] 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)
- [17] Jia, S., Qian, Y.T.: Constrained Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* 47, 161–173 (2009)
- [18] Liu, X.S., Xia, W., Wang, B., Zhang, L.M.: An Approach Based on Constrained Nonnegative Matrix Factorization to Unmix Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* 49, 757–772 (2011)
- [19] Bayliss, J., Gaultieri, J.A., Cromp, R.F.: Analyzing hyperspectral data with independent component analysis. In: Proc. SPIE, vol. 3240, pp. 133–143 (1997)
- [20] Chen, C., Zhang, X.: Independent component analysis for remote sensing study. In: Proc. SPIE, vol. 3871, pp. 150–158 (1999)
- [21] Moussaoui, S., Hauksdottir, H., Schmidt, F., Jutten, C., Chanussot, J., Brie, D., Doute, S., Benediktsson, J.A.: On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation. *Neurocomputing* 71(10-12), 2194–2208 (2008)
- [22] Nascimento, J., Bioucas-Dias, J.: Hyperspectral unmixing algorithm via dependent component analysis. In: Proc. IEEE IGARSS, pp. 4033–4036 (July 2007)
- [23] Xia, W., Liu, X., Wang, B., Zhang, L.: Independent Component Analysis for Blind Unmixing of Hyperspectral Imagery with Additional Constraints. *IEEE Transactions on Geoscience and Remote Sensing* 49(6), 2165–2179 (2011)

GPCA on Gabor Tensor for Face Recognition

Lian Zhu¹, Rong Huang¹, Xinfu Ye¹, Wankou Yang^{1,2}, and Sun Changyin¹

¹ School of Automation, Southeast University, Nanjing 210096, P.R. China

² Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing 210094, P.R. China
zhulianseu@163.com

Abstract. There is a growing interest in subspace learning techniques for face recognition such as Gabor face representation. Although the Gabor face representation has received great success in face recognition, the excessive dimension of the data space often brings the algorithms into the curse of dimensionality dilemma. This paper proposes a novel face recognition method based on discriminant analysis with Gabor tensor representation. We derive a 3rd-order Gabor tensor representation from a complete response set of 40 Gabor filters. Then Generalized Principal Component Analysis (GPCA) is applied to each Gabor feature matrix. After working out 40 times, the feature matrices in a lower dimensional subspace are finally integrated for classification. Experimental results on ORL database and AR database show promising results of the proposed method.

Keywords: Gabor tensor representation, GPCA, face recognition.

1 Introduction

In the real world, the extracted feature of an object often has some specialized structures and such structures are in the form of second or even higher order tensors [1]. For example, gray-level images indicate second order tensor data (a matrix), and can be expanded to a third order tensor with the inclusion of temporal data, such as video sequences in content analysis [2], or by representing sets of Gabor filter images [3], [4] as often used in face recognition.

In this paper, each image is represented by a Gabor tensor which is composed of 40 Gabor feature matrices. There are three major reasons for introducing the Gabor-based representation: 1) human brains seem to have a special function to process information in multiresolution levels [5], [6], [7], which can be simulated by controlling the scale parameter in Gabor functions; 2) it is supposed that Gabor functions are similar to the receptive field profiles in the mammalian cortical simple cells , and 3) Gabor-function-based representations have been successfully employed in many computer vision applications such as face recognition[8], [9], object recognition, and texture analysis [10]. However, Gabor features are usually very high-dimensional data and there are redundancies among them. Most traditional algorithms, such as the Principal Components Analysis (PCA) [11], input an image object as a 1-D vector. The dimensionality of the concatenated 1-D data in PCA is usually very high, while in real object recognition applications such as face recognition the available number of

training samples is small. In consequence, subspaces learnt from PCA may inadequately represent a sample distribution, resulting in poor recognition performance.

To avoid this, it is often helpful to process the data in its original form and order. In [12], Yang et al. proposed a scheme called 2DPCA to conduct dimensionality reduction with a 2-D matrix representation. We apply the Generalized Principal Component Analysis (GPCA) for the dimension reduction of Gabor feature. Experiments on face databases show that, the proposed GPCA on Gabor Tensor algorithm outperforms the traditional vector-based subspace learning algorithms.

2 Gabor Tensor Representation

The representation of faces using Gabor features has been extensively and successfully used in face recognition. The family of 2D Gabor kernel filters composed by five frequencies and eight orientations can be defined as follows:

$$\psi_{\omega,v} = \frac{k_{\omega,v}^2}{\sigma^2} \exp\left(-\frac{k_{\omega,v}^2 z^2}{2\sigma^2}\right) [\exp(-ik_{\omega,v}z) - \exp(-\frac{\sigma^2}{2})] \quad (1)$$

where ω and v define the orientation and scale of the Gabor kernels respectively, $z = (x, y)$, and the wave vector $k_{\omega,v}$ is defined as follows:

$$k_{\omega,v} = k_v e^{i\phi_\omega} \quad (2)$$

where $k_v = k_{\max} / f^v$, $k_{\max} = \pi / 2$, $f = \sqrt{2}$, $\phi_\omega = \pi\omega / 8$.

In this paper, we use Gabor kernels at five scales $v \in \{0, 1, 2, 3, 4\}$ and eight orientations $\omega \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ to derive the Gabor representation. The response of an image $I(x, y)$ to a wavelet $\psi_{\omega,v}(z)$ is obtained by the convolution:

$$G_{\omega,v}(x, y) = I(x, y) * \psi_{\omega,v}(z) \quad (3)$$

The Gabor wavelet coefficient obtained for a given scale and orientation in Equation(3), is a complex number. It has been discovered that the magnitude varies slowly, while the phase information varies its rotation with the spatial position. For this reason, only the magnitude is usually used for face classification.



Fig. 1. A face image and its corresponding 3rd-order Gabor tensor

For every image pixel we have totally 40 Gabor magnitude coefficients which can be regarded as a Gabor feature vector of 40 dimensions. Therefore, a $h \times w$ 2D image can be encoded by 40 Gabor filters to form a $40 \times h \times w$ 3rd-order Gabor tensor. Fig. 1 shows an example of a face image with its corresponding 3rd-order Gabor tensor.

3 Generalized Principal Component Analysis (GPCA)

3.1 Principal Component Analysis

Image can be represented by a matrix $X \in R^{r \times c}$, where r and c are the number of rows and columns in the image. Each image matrix X_i can be vectorized to a vector x_i by concatenating all the rows in X_i . The PCA transformation that preserves the principal components is given by:

$$Y = XW \quad (4)$$

PCA is to obtain a k -dimensional feature projection matrix, which converts the feature vectors from the high-dimensional to low-dimensional.

3.2 GPCA

The key difference between PCA and GPCA is in the representation of image data. While PCA uses a vectorized representation of the 2D image matrix, GPCA works with a 2D matrix representation (as illustrated schematically in Fig. 2) and attempts to preserve spatial locality of the pixels.

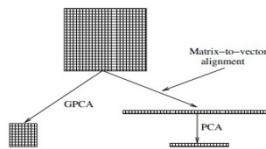


Fig. 2. Key difference between PCA and GPCA

In GPCA, we consider images as two dimensional signals and we define the following (l_1, l_2) -dimensional axis space: $u_i \otimes v_j$ for $i=1, \dots, l_1$ and $j=1, \dots, l_2$, where \otimes denotes the tensor product, $u_i \in R^{r \times l_1}, v_j \in R^{c \times l_2}$. For a given matrix $X \in R^{r \times c}$, its projection onto the (i, j) -th coordinate is $u_i \cdot X \cdot v_j$. In GPCA, we search for an optimal (l_1, l_2) -dimensional axis space $u_i \otimes v_j$, for $i=1, \dots, l_1$ and $j=1, \dots, l_2$, such that the projections of the data points onto this axis space have the maximum variance. Unlike PCA, however, the projections of the data points onto the (l_1, l_2) -dimensional axis system in GPCA are matrices, instead of vectors.

The mean and variance of a set of image matrices can be defined as follows:

Definition. Let $S = \{X_1, \dots, X_n\}$ be a set of matrices in $R^{r \times c}$. Then the variance of S is defined as $\text{var}(S) = \frac{1}{n-1} \sum_{i=1}^n \|X_i - \bar{X}\|_F^2$, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is the mean of S , and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

A simple way to compute the projection is to form two matrices $L = (u_1, \dots, u_{l_1})$ and $R = (v_1, \dots, v_{l_2})$. The projection can then be computed by $L^T X R \in R^{l_1 \times l_2}$. Let $A_i \in R^{r \times c}$ for

$i=1,\dots,n$ be the n images in the dataset and $M = \frac{1}{n} \sum_{i=1}^n A_i$ be their mean. Let $\tilde{A}_i = A_i - M$, for all i . Then the variance of the projection of $\{\tilde{A}_i\}_{i=1}^n$ onto the (l_1, l_2) -dimensional axis system can be computed as

$$\text{var}(L, R) = \frac{1}{n-1} \sum_{i=1}^n \|L^T \tilde{A}_i R\|_F^2 \quad (5)$$

where $L = [u_1, \dots, u_{l_1}] \in R^{r \cdot l_1}$ and $R = [v_1, \dots, v_{l_2}] \in R^{c \cdot l_2}$.

GPCA aims to compute two matrices $L \in R^{r \cdot l_1}$ and $R \in R^{c \cdot l_2}$ with orthonormal columns, such that the variance $\text{var}(L, R)$ is maximized. Considering $\text{var}(L, R) = \frac{1}{n-1} \sum_{i=1}^n \|L^T \tilde{A}_i R\|_F^2$, for a given R , matrix L consists of the l_1 eigenvectors of the matrix $M_L = \sum_{i=1}^n \tilde{A}_i R R^T \tilde{A}_i^T$ corresponding to the largest l_1 eigenvalues for a given L , matrix R consists of the l_2 eigenvectors of the matrix $M_R = \sum_{i=1}^n \tilde{A}_i^T L L^T \tilde{A}_i$ corresponding to the largest l_2 eigenvalues.

When reducing the dimension using GPCA, we take $l_1 = l_2 = d$ for simplicity. The analysis above provides us an iterative algorithm for computing L and R . We define the root mean square error (RMSE), to measure the average reconstruction error for \tilde{A}_i as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \|\tilde{A}_i - LL^T \tilde{A}_i R R^T\|_F^2} \quad (6)$$

The pseudo-code for the GPCA algorithm is could be found in [13].

4 GPCA on Gabor Tensor and Experiment Result

4.1 GPCA on Gabor Tensor

In this part, we propose an idea which applies the GPCA to Gabor tensor. First, we compute the Gabor features $G_{i1}, G_{i2}, \dots, G_{i40}$ for each image sample which can form a Gabor Tensor cube T_i , where i stands for the $i-th$ sample. Then considering every Gabor feature, we employ the GPCA method to the training set of N training samples $\{G_{1j}, G_{2j}, \dots, G_{Nj}\}$ and work out the matrix L_j and R_j for the dimension reduction, where j refers to the $j-th$ Gabor feature. Next, we utilize the L_j and R_j to reduce Gabor feature matrix dimension of the training samples and test samples. At last, we compute the distance between the dimension reducted feature cubes and decide the class the test samples belong to. The procedure is displayed in fig. 3.

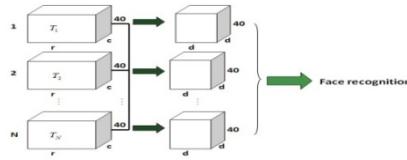


Fig. 3. Procedure of our method

4.2 Experiment Result

We apply our idea in the face recognition and test it on the ORL and AR database. The ORL face database contains 10 different images of 40 distinct subjects. The images are grey scale with a resolution of 92×112 . Fig. 4 gives the ten different images of a person in the ORL database.



Fig. 4. 10 images of one person in the ORL database

The AR face database contains over 4,000 color images corresponding to 126 people's faces. The database in our experiment contains 14 different images of 120 distinct subjects and the images are grey scale with a resolution of 40×50 . The images of the first subject are showed below in Fig. 5.



Fig. 5. 14 images of one person in the AR database

The following content will show the performance of our method against some traditional methods. The recognition rate on AR database is plotted in fig. 6.

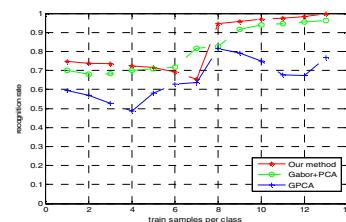


Fig. 6. Face recognition rate on AR database

The results on two databases demonstrate that our method show better performance than the Gabor+PCA and GPCA.

4.3 Parameters

There are mainly two parameters that affect the performance of our method: the Gabor kernel size $s \times s$ and the dimension d of GPCA. In order to study the effect of the parameters, we conduct a series of experiments on the YALE face database. The rate along with the change of the parameters is plotted in the following Fig. 7.

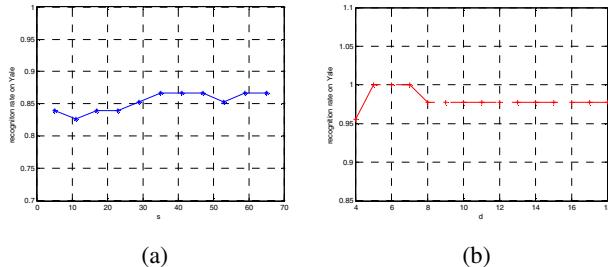


Fig. 7. (a) The effect of Gabor kernel side length s ; (b) The influence of GPCA dimension d

5 Conclusions

In our method, the image objects were encoded as an 3rd-order gabor tensor. The GPCA is presented to project the high-dimensional data into low-dimensional space. Compared with traditional algorithms, such as Gabor feature and PCA, our algorithm effectively avoids the curse of dimensionality dilemma and alleviates the small sample size problem.

Acknowledgments. This work is supported in part by NSF of China (61005008, 61375001) and in part by the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology) (309201130122006).

References

1. Vasilescu, M., Terzopoulos, D.: Multilinear subspace analysis for image ensembles. In: Proc. Computer Vision and Pattern Recognition, Madison, WI, vol. 2, pp. 93–99 (June 2003)
2. Dimitrova, N., Zhang, H., Shahraray, B., Sezan, I., Huang, T., Za-khor, A.: Applications of video-content analysis and retrieval. Proc. IEEE Multimedia 9(3), 42–55 (2002)
3. Hamamoto, H., Uchimura, S., Watanabe, M., Yasuda, T., Tomita, S.: Recognition of handwriting numerals using gabor features. In: Proc. IEEE Conf. Pattern Recognit, pp. 250–253 (1996)
4. Wiskott, L., Fellous, J.M., Kruger, N., Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 775–779 (1997)
5. Marcelja, S.: Mathematical Description of the Responses of Simple Cortical Cells. J. Optical Soc. Am. 70(11), 1297–1300 (1980)

6. Daugman, J.G.: Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile. *Vision Research* 20, 847–856 (1980)
7. Daugman, J.G.: Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters. *J. Optical Soc. Am.* 2(7), 1160–1169 (1985)
8. Liu, C., Wechsler, H.: Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition. *IEEE Trans. Image Processing* 11(4), 467–476 (2002)
9. Liu, C.: Gabor-Based Kernel PCA with Fractional Power Polynomial Models for Face Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 26(5), 572–581 (2004)
10. Dunn, D., Higgins, W.E., Wakeley, J.: Texture Segmentation Using 2D Gabor Elementary Functions. *IEEE Trans. Pattern Analysis and Machine Intelligence* 16(2), 130–149 (1994)
11. Turk, M., Pentland, A.: Eigenfaces for recognition. *J. Cognitive Neurosci.* 3(1), 71–86 (1991)
12. Yang, J., Zhang, D., Frangi, A., Yang, J.: Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 26(1), 131–137 (2004)
13. Ye, J., Janardan, R., Li, Q.: GPCA: an efficient dimension reduction scheme for image compression and retrieval. In: *KDD 2004*, pp. 354–363 (2004)

Nonnegative Discriminative Manifold Learning for Hyperspectral Data Dimension Reduction

Lefei Zhang¹, Liangpei Zhang², Dacheng Tao³, Xin Huang², and Bo Du¹

¹ Computer School, Wuhan University

² LIESMARS, Wuhan University

³ Centre for Quantum Computation and Intelligent Systems

University of Technology, Sydney

gospace@163.com

Abstract. Manifold learning algorithms have been demonstrated to be effective for hyperspectral data dimension reduction (DR). However, the low dimensional feature representation resulted by traditional manifold learning algorithms could not preserve the nonnegative property of the hyperspectral data, which leads inconsistency with the psychological intuition of “combining parts to form a whole”. In this paper, we introduce a nonnegative discriminative manifold learning (NDML) algorithm for hyperspectral data DR, which yields a discriminative and low dimensional feature representation, with psychological and physical evidence in the human brain. Our method benefits from both the non-negative matrix factorization (NMF) algorithm and the discriminative manifold learning (DML) algorithm. We apply the NDML algorithm to hyperspectral remote sensing image classification on HYDICE dataset. Experimental results confirm the efficiency of the proposed NDML algorithm, compared with some existing manifold learning based DR methods.

Keywords: manifold learning, nonnegative matrix factorization, dimension reduction, hyperspectral data.

1 Introduction

Dimension reduction (DR) plays an important role in hyperspectral remote sensing image (HRSI) classification because it can (1) reduce the redundancy among the input features, (2) decrease the computational cost in subsequent processing, and (3) preserve the discriminative information that benefits for classification [1-3]. Manifold learning algorithms [4, 5], which aim to find a certain feature mapping from the original high dimensional feature to the reduced feature representation, have been demonstrated to be effective in hyperspectral data DR [6, 7]. Among them, discriminative manifold learning (DML) algorithms have significantly showed their outstanding performance in HRSI classification, for instance, generalised supervised local tangent space alignment (GSLTSA) [8], local Fisher’ s discriminant analysis (LFDA) [9], discriminative metric learning [10] and semi-supervised discriminative locally enhanced alignment (SDLEA) [11], etc.

However, in the manifold learning algorithms mentioned above, the resulted low dimensional feature representation could not preserve the nonnegativity of the input hyperspectral data, which leads inconsistency with the psychological intuition of combining parts to form a whole [12]. Nonnegative matrix factorization (NMF) [13] is one of the most important matrix factorization techniques that have been frequently applied in machine learning and pattern recognition area [14]. NMF finds two non-negative matrices (called bases and coefficient, respectively) whose product provides a good approximation to the observed feature matrix. The nonnegative constraints in NMF consequently lead to a parts-based representation, since only additive (not subtractive) combinations are allowed.

This paper is inspired both by the discriminative power of the DML and by the nonnegative property of NMF. More precisely, we introduce a nonnegative discriminative manifold learning (NDML) algorithm, which combines the NMF and DML together to obtain a nonnegative and discriminative low dimensional feature representation with the psychological and physical explanation. NDML algorithm could be utilized for hyperspectral data dimension reduction. The subsequent image classification results confirm the improvement of classification overall accuracy (OA).

The rest of this paper is outlined as follows. Section 2 first presents the proposed NDML algorithm; Section 3 then shows the HRSI classification results on HYDICE dataset; and Section 4 finally concludes the paper.

2 Nonnegative Discriminative Manifold Learning

Suppose the input feature matrix of NDML (spectral feature matrix of training samples in hyperspectral data) is $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N] \in \mathbb{R}^{P \times N}$ and the related ground truth label of each sample $\mathbf{v}_i \in \mathbb{R}^P$ to be $l_i = \{1, 2, \dots, c\}$ ($i=1, 2, \dots, N$), in which N , P and c denote the number of samples, features and classes, respectively. NDML algorithm aims to find a low dimensional feature representation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] \in \mathbb{R}^{d \times N}$ in which the nonnegative property of hyperspectral data as well as the discriminative information from training samples could be well preserved. Based on patch alignment framework [15], for each sample \mathbf{v}_i , we build two local patches using (1) the k_1 nearest samples that with same label of \mathbf{v}_i , and (2) the k_2 nearest samples that with different label of \mathbf{v}_i , i.e.,

$$\mathbf{V}_i^s = [\mathbf{v}_i, \mathbf{v}_1^s, \mathbf{v}_2^s, \dots, \mathbf{v}_{k_1}^s] \in \mathbb{R}^{P \times (k_1+1)} \quad (1)$$

$$\mathbf{V}_i^d = [\mathbf{v}_i, \mathbf{v}_1^d, \mathbf{v}_2^d, \dots, \mathbf{v}_{k_2}^d] \in \mathbb{R}^{P \times (k_2+1)} \quad (2)$$

Based on the supervised information, the discriminative local geometric structure could be preserved by minimize the distances of sample pairs of the same class as well as maximize the distances of sample pairs of the different classes in the low dimensional feature space. By denote \mathbf{H}_i^s and \mathbf{H}_i^d as the low dimensional feature representation of \mathbf{V}_i^s and \mathbf{V}_i^d , we have the following patch optimizations:

$$\min_{\mathbf{H}_i^s} \text{tr}(\mathbf{H}_i^s \mathbf{L}_i^s \mathbf{H}_i^{sT}), \quad \max_{\mathbf{H}_i^d} \text{tr}(\mathbf{H}_i^d \mathbf{L}_i^d \mathbf{H}_i^{dT}) \quad (3)$$

in which,

$$\mathbf{L}_i^s = \begin{bmatrix} k_1 & -\mathbf{e}_{k_1}^T \\ -\mathbf{e}_{k_1} & \mathbf{I}_{k_1} \end{bmatrix}, \quad \mathbf{L}_i^d = \begin{bmatrix} k_2 & -\mathbf{e}_{k_2}^T \\ -\mathbf{e}_{k_2} & \mathbf{I}_{k_2} \end{bmatrix} \quad (4)$$

in which $\mathbf{e}_k = [1, \dots, 1]^T \in \mathbb{R}^k$ and $\mathbf{I}_k \in \mathbb{R}^{k \times k}$ is an identity matrix.

Then the full optimization of discriminative manifold learning is obtained by summing all the patch optimizations of \mathbf{v}_i :

$$\min_{\mathbf{H}} \sum_{i=1}^N \text{tr}(\mathbf{H}_i^s \mathbf{L}_i^s \mathbf{H}_i^{sT}) = \min_{\mathbf{H}} \text{tr}(\mathbf{HL}^s \mathbf{H}^T) \quad (5)$$

$$\max_{\mathbf{H}} \sum_{i=1}^N \text{tr}(\mathbf{H}_i^d \mathbf{L}_i^d \mathbf{H}_i^{dT}) = \max_{\mathbf{H}} \text{tr}(\mathbf{HL}^d \mathbf{H}^T) \quad (6)$$

where $\mathbf{L}^s = \sum_{i=1}^N \mathbf{S}_i^s \mathbf{L}_i^s \mathbf{S}_i^{sT}$ and $\mathbf{L}^d = \sum_{i=1}^N \mathbf{S}_i^d \mathbf{L}_i^d \mathbf{S}_i^{dT}$ are the alignment matrices of same class distance minimization and different class distance maximization, respectively. \mathbf{S}_i^s and \mathbf{S}_i^d are selection matrices to identify the global index of each sample in \mathbf{H}_i^s and \mathbf{H}_i^d . Since both \mathbf{L}^s and \mathbf{L}^d are symmetric and positive semi-definite, we combining (5) and (6) to:

$$\min_{\mathbf{H}} \text{tr}(\mathbf{H} (\mathbf{L}^{d-1/2})^T \mathbf{L}^s \mathbf{L}^{d-1/2} \mathbf{H}^T) = \min_{\mathbf{H}} \text{tr}(\mathbf{HLH}^T) \quad (7)$$

Eq. (7) is the standard form of manifold learning DR algorithms [15, 16], however, the solution of (7) could not maintain the nonnegative property of input hyperspectral data. In this paper, we consider to combine the NMF and DML together. From the perspective of DR, NMF aims to represent input feature matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N] \in \mathbb{R}^{P \times N}$ as a linear combination of low dimensional bases $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d] \in \mathbb{R}^{P \times d}$ and coefficient matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] \in \mathbb{R}^{d \times N}$, in which both bases and coefficient are nonnegative. The optimization of NMF is:

$$\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} D(\mathbf{V}, \mathbf{WH}) \quad (8)$$

in which $D(\mathbf{V}, \mathbf{WH})$ denotes the distance of the input feature matrix \mathbf{V} and its low rank approximation \mathbf{WH} , which could be measured by the conventional least squares error or the generalized K-L divergence [17]. Here we use the latter, which is defined as:

$$D(\mathbf{V}, \mathbf{WH}) = \sum_{i=1}^P \sum_{j=1}^N \left(V_{ij} \log \frac{V_{ij}}{(\mathbf{WH})_{ij}} - V_{ij} + (\mathbf{WH})_{ij} \right) \quad (9)$$

In (8), the coefficient matrix \mathbf{H} is the output low dimensional feature representation of \mathbf{V} . Then, by adding (8) as the nonnegative constraint of DML, we have the final objective optimization of NDML:

$$\min_{W \geq 0, H \geq 0} \text{tr}(HLH^T) + \beta D(V, WH) \quad (10)$$

where $\beta > 0$ is a trade-off parameter between DML and its nonnegative constraint.

Note that (10) is non-convex on both W and H , so it is impossible to find the global optimal solution of NDML. In the literature, multiplicative update rules (MUR) [17, 18] is a good method for solving NMF algorithms from both speed and ease of implementation points of view. Thus, we apply MUR to solve (10). However, the MUR converges slowly because it is fundamental a first-order method. In our previous work, a new efficient fast gradient descent (FGD) was proposed to overcome the slow convergence of MUR, which leads the solution of (10). Further detailed information could refer to [12].

Similar to DR algorithms [7, 19], the NDML suffers from the out-of-sample problem, because the feature mapping from V to H is nonlinear and implicit. In order to overcome this issue, we propose the linear version of NDML. For all the training samples, since V can be approximated by WH , we can then project an arbitrary test sample from the original feature $x \in R^P$ to the reduced low dimensional $y \in R^d$ by the following linear transformation using the pseudo-inverse of W :

$$y = (W^T W)^{-1} W^T x \quad (11)$$

3 Experimental Results

The experimental analysis is conducted on a public hyperspectral remote sensing image provided by Purdue University [20]. This dataset is an urban site of the airborne hyperspectral digital imagery collection experiment (HYDICE) from the mall in Washington, DC, which has an original size of 1280×307 pixels. A total of 210 bands are collected in the $0.4\text{--}2.4 \mu\text{m}$ region of the visible and infrared spectra. The water absorption bands are then deleted, resulting in 191 channels. In this study, we use a subset of the whole set, with a size of 280×307 pixels, as shown in Fig. 1 (a).

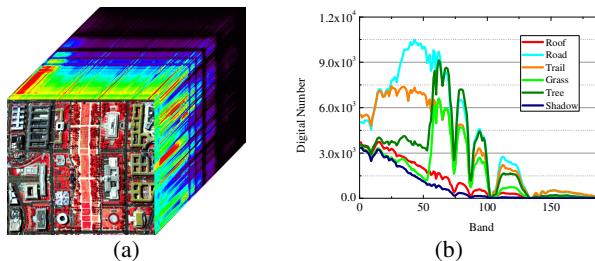


Fig. 1. (a) RGB composites of HYDICE dataset (bands 65, 52, and 36 for red, green, and blue, respectively), (b) representative spectral curves in HYDICE dataset.

There are numerous of studies focus on this dataset and it is challenging to analyze because some of the land cover classes are spectrally similar [21-23]. The desired information classes in this dataset are: roof, road, trail, grass, tree and shadow, respectively. The representative spectral curves of aforementioned classes are plotted in

Fig. 1 (b), Fig 2 shows the ground truth map of the reference data, the numbers of training and test samples for classification are listed in Table 1.

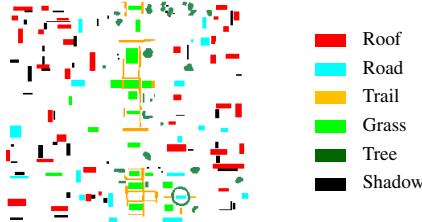


Fig. 2. Reference data of HYDICE dataset

Table 1. Numbers of all reference data, training and test samples for the HYDICE data set

	All reference data	Training samples	Test samples
Roof	3129	50	3079
Road	1402	50	1352
Trail	1267	50	1217
Grass	1790	50	1740
Tree	1194	50	1144
Shadow	1120	50	1070
Total	9902	300	9602

The classification maps of HYDICE image using different features are shown in Fig. 3, based on support vector machine (SVM) [24, 25]. To compare the effectiveness of the proposed NDML with the conventional manifold learning DR methods, we show the performance of Locality Preserving Projection (LPP) [26] and Neighborhood Preserving Embedding (NPE) [27]; in these feature DR algorithms, we fix the subspace feature dimensionality as $d=50$. We also address the classification result of original spectral feature as the baseline for DR methods. The detailed class-specific rates in percentage are reported in Table 2. It could be observed that the discussed NDML algorithm achieves the best classification results in both accuracy and visual interpretation.

Table 2. Class-specific rates in percentage for various features in HYDICE dataset

	Original	LPP	NPE	NDML
Roof	84.67	83.63	87.20	91.16
Road	96.37	83.87	92.15	96.44
Trail	97.12	95.89	96.22	97.12
Grass	98.21	97.70	97.47	98.96
Tree	93.88	96.91	96.72	97.10
Shadow	98.41	92.74	95.80	97.02
OA	92.98	90.34	92.99	95.44
Kappa	0.9133	0.8802	0.9131	0.9434

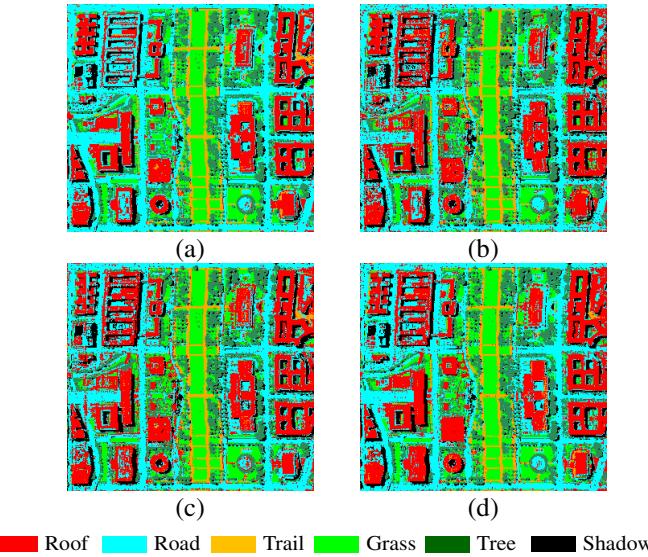


Fig. 3. Classification maps of all the methods in HYDICE dataset, based on SVM. (a) original, (b) LPP, (c) NPE, (d) NDML.

For a more detailed comparison of the performance of these DR algorithms, the classifications are conducted using the aforementioned DR algorithms with an increase in subspace feature dimensionality d . Fig. 4 shows the classification rates under the three algorithms. As shown in Fig. 4, the NDML performs better than the other two algorithms when $d > 20$ and achieves the best performance around $d = 60$.

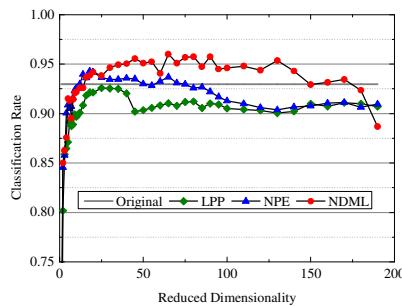


Fig. 4. Relationship of subspace feature dimensionality and classification rate in HYDICE dataset for DR approaches

4 Conclusion

This manuscript proposes a novel nonnegative discriminative manifold learning (NDML) algorithm for hyperspectral remote sensing image (HRSI) dimension reduction. By combining the nonnegative matrix factorization (NMF) and discriminative

manifold learning (DML) together, NDML aims to achieve a discriminative low dimensional feature representation with the psychological and physical evidence of the parts based representation in the human brain. Experimental results on HYDICE image classification prove that the NDML algorithm outperforms some state-of-the-art manifold learning dimension reduction methods.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under Grants 61102128, and the National Basic Research Program of China (973 Program) under Grant 2012CB719905.

References

1. Harsanyi, J.C., Chang, C.-I.: Hyperspectral Image Classification and Dimensionality Reduction: An Orthogonal Subspace Projection Approach. *IEEE Trans. Geosci. Remote Sens.* 32(4), 779–785 (1994)
2. Jimenez, L.O., Landgrebe, D.A.: Hyperspectral Data Analysis and Supervised Feature Reduction Via Projection Pursuit. *IEEE Trans. Geosci. Remote Sens.* 37(6), 2653–2667 (1999)
3. Zhang, L., Zhang, L., Tao, D., Huang, X.: Tensor Discriminative Locality Alignment for Hyperspectral Image Spectral-Spatial Feature Extraction. *IEEE Trans. Geosci. Remote Sens.* 51(1), 242–256 (2013)
4. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(22), 2323–2326 (2000)
5. Balasubramanian, M., Schwartz, E.L.: The Isomap Algorithm and Topological Stability. *Science* 295(5552), 7 (2002)
6. Bachmann, C.M., Ainsworth, T.L., Fusina, R.A.: Exploiting Manifold Geometry in Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* 43(3), 441–454 (2005)
7. Zhang, L., Zhang, L., Tao, D., Huang, X.: On Combining Multiple Features for Hyperspectral Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* 50(3), 879–893 (2012)
8. Ma, L., Crawford, M.M., Tian, J.: Generalised Supervised Local Tangent Space Alignment for Hyperspectral Image Classification. *Electron. Lett.* 46(7), 497–498 (2010)
9. Li, W., Prasad, S., Fowler, J.E., Bruce, L.M.: Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis. *IEEE Trans. Geosci. Remote Sens.* 50(4), 1185–1198 (2012)
10. Du, B., Zhang, L., Zhang, L., Chen, T., Wu, K.: A Discriminative Manifold Learning Based Dimension Reduction Method. *Int. J. Fuzzy Syst.* 14(2), 272–277 (2012)
11. Shi, Q., Zhang, L., Du, B.: Semi-Supervised Discriminative Locally Enhanced Alignment for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* 51 (2013) doi:10.1109/TGRS.2012.2230445
12. Guan, N., Tao, D., Luo, Z., Yuan, B.: Non-Negative Patch Alignment Framework. *IEEE Trans. Neural Netw.* 22(8), 1218–1230 (2011)
13. Lee, D.D., Seung, H.S.: Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* 401(6755), 788–791 (1999)
14. Cai, D., He, X., Han, J., Huang, T.S.: Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(8), 1548–1560 (2011)

15. Zhang, T., Tao, D., Li, X., Yang, J.: Patch Alignment for Dimensionality Reduction. *IEEE Trans. Knowl. Data Eng.* 21(9), 1299–1313 (2009)
16. Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S.: Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(1), 40–51 (2007)
17. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. In: NIPS, vol. 13, pp. 556–562 (2001)
18. Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: An Optimal Gradient Method for Non-negative Matrix Factorization. *IEEE Trans. Signal Process.* 60(6), 2882–2898 (2012)
19. Shi, L., Zhang, L., Yang, J., Zhang, L., Li, P.: Supervised Graph Embedding for Polarimetric SAR Image Classification. *IEEE Geosci. Remote Sens. Lett.* 10(2), 216–220 (2013)
20. <https://engineering.purdue.edu/~biehl/MultiSpec/>
21. Benediktsson, J.A., Palmason, J.A., Sveinsson, J.R.: Classification of Hyperspectral Data From Urban Areas Based on Extended Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* 43(3), 480–491 (2005)
22. Li, C.-H., Kuo, B.-C., Lin, C.-T., Huang, C.-S.: A Spatial-Contextual Support Vector Machine for Remotely Sensed Image Classification. *IEEE Trans. Geosci. Remote Sens.* 50(3), 784–799 (2012)
23. Huang, X., Zhang, L.: An Adaptive Mean-Shift Analysis Approach for Object Extraction and Classification From Urban Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sens.* 46(12), 4173–4185 (2008)
24. Mountrakis, G., Im, J., Ogole, C.: Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm.* 66(3), 247–259 (2011)
25. Bazi, Y., Melgani, F.: Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* 44(11), 3374–3385 (2006)
26. He, X., Niyogi, P.: Locality Preserving Projections. In: NIPS, vol. 16, pp. 153–160 (2004)
27. He, X., Cai, D., Yan, S., Zhang, H.-J.: Neighborhood Preserving Embedding. In: ICCV, vol. 2, pp. 1208–1213 (2005)

Spectral Unmixing for Hyperspectral Image Classification with an Adaptive Endmember Selection

Qingjie Meng, Yanning Zhang, Wei Wei, and Lei Zhang

School of Computer Science, Shaanxi Province Key Laboratory of Speech & Image Information Processing, Northwestern Polytechnical University, Xi'an, China

mqjcatherine@163.com, ynzhang@nwpu.edu.cn

weiweinnwpu@nwpu.edu.com, zhanglei_wonderful@126.com

Abstract. Hyperspectral classification techniques are widely used for detailed analysis of the earth surface. However, mixed pixels caused by the relatively low spatial resolution of the imaging system are the big burden for traditional pure-pixel-hypothesis based hard classification methods. To address this problem, a novel method, which jointly uses soft classification and spectral unmixing, is proposed in this paper. The confusion matrix is exploited to determine the endmember set for each class. Then the generated endmember is adopted for spectral unmixing. The fractional abundance of training samples, which is generated from spectral unmixing, is utilized to optimize soft multinomial logistic regression classifier. The result of the optimized classifier will result in a more accurate confusion matrix. Thus, this procedure is executed iteratively to achieve required performance. Experimental results on synthetic and real hyperspectral data sets demonstrate the superiority of the proposed method for hyperspectral image classification.

Keywords: hyperspectral image, spectral unmixing, supervised classification, endmember selection.

1 Introduction

Hyperspectral image is characterized by a high spectral resolution at the micron level and a relatively low spatial resolution at the meter level. Image classification [1] and spectral unmixing [2] are two significant techniques for remote sensing data analysis. The abundant wavelength information provided by a very high spectral resolution allows detailed classification and detection of land cover type. However, mixed pixels are resulted by relatively low spatial resolution since several pure spectral signatures (or endmembers) are overlapped at sub-pixel level. Traditional hard classification methods are not suitable for scenarios with mixed pixels [3]. As complementary methods, hard classification and spectral unmixing are combined to offer a promising direction.

Although jointly using of classification and spectral unmixing algorithms exhibits potential superiority to address mixed pixel problem for hyperspectral image classification, it is rarely investigated so far [3] [4]. As an early attempt, an unsupervised

approach for mixed pixel classification in hyperspectral image is proposed [5]. It achieved finer results compared with widely used traditional classification methods. However, the drawback of unsupervised method is appeared on sophisticated dataset. Then a supervised technique has been introduced [3]. Soft classification (probabilistic SVM) is explored to get preliminary classification map while spectral unmixing is used to solve sub-pixel mixing problem in the rough map. In synthetic and real data set, land cover maps with an improved spatial resolution are obtained. Recently, a semi-supervised algorithm, which integrates a well established discriminative classifier (multinomial logistic regression [6]) with linear spectral unmixing, is proposed in [4]. Both limited labeled and unlabeled training samples selected by active learning approach are utilized to train parameters of multinomial logistic regression (MLR) [7]. In this method, soft labels of unlabeled samples are provided by spectral unmixing. Results on synthetic and AVIRIS data validate effectiveness of the method compared with probabilistic SVM classification.

In this paper, a novel classification method, which jointly uses spectral unmixing technique, is proposed to handle mixed pixel problem for hyperspectral image. Endmember set of a certain class is selected via confusion matrix. Then the generated endmember set is adopted for the fully constrained least squares (FCLS) [8] unmixing model. The probabilistic labels of training samples, which are achieved by FCLS unmixing, are used to optimize soft multinomial logistic regression (MLR) classifier. Thus a more reliable confusion matrix is obtained by the optimized classifier. This procedure is executed iteratively to reach required performance. Experiments on the simulated and real hyperspectral data sets indicate the effectiveness of the proposed method trained on randomly chosen sample set and the superiority of proposed method over commonly used traditional methods and probabilistic SVM [3].

The paper is organized as follows: the proposed approach is described in detail in Section 2. Experiment on synthetic and real hyperspectral data sets is designed in Section 3. Conclusion is drawn in Section 4.

2 Adaptive Endmember Selection for Classification

The proposed method in this paper is an iteration process mainly including three steps as Fig. 1 demonstrates. Step 1, endmember set of each class is determined by the confusion matrix of classification outputted from last cycle. This selection based on an assumption that classes (m_j, m_k), to which a relatively larger number of pixels belonging m_i are wrongly classified after a reliable classifier, have more probability to be confused with class i. Step 2, the linear spectral unmixing is performed on training samples according to above endmember set per class. Fully constrained least square (FCLS) method is applied to assess fractional abundance of every training sample. Step 3, these obtained fractional abundance, which are taken as soft labels, are exploited to optimize a soft classifier. In this paper, multinomial logistic regression (MLR) discriminative classification is adopted due to its ability to process large data size effectively and to produce sparse result [7].

The proposed adaptive endmember selection mechanism is introduced in detail in subsection 2.1. Soft MLR-based classification method is described in subsection 2.2. The outline of proposed method is in subsection 2.3.

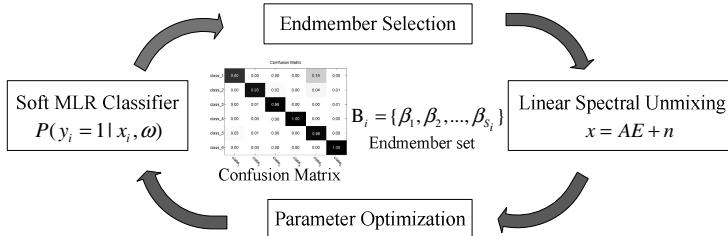


Fig. 1. Flow chat of the proposed iterative method

2.1 Adaptive Endmember Selection Model

In this stage, the proposed endmember selection model is established to determine possible land cover types of each class by taking advantages of confusion matrix.

Due to the fact that confusion matrix $C \in \mathbb{R}^{M \times M}$ allows visualization of mislabeling or confusing between classes, it is reliable to determine overlapped classes. Endmember selection model is presented in Eq.(1) and Eq.(2).

$$B_i = E_T, i \in M \quad (1)$$

$$T = \{t \mid C_{it} \neq 0, C_{it} \geq \lambda C_{ii}, t \in M\} \quad (2)$$

where $M = [1, 2, \dots, M]$ describes a set of M class labels. $E = [e_1, e_2, \dots, e_M]$ denotes endmember set of the image. $B_i = [\beta_1, \beta_2, \dots, \beta_{S_i}]$ is the selected endmember set of class i . E_T represents endmember subset chosen from E . Since land cover types possibly included in each class are just a small subset of all classes in the image, numbers of endmembers in each class S_i satisfy $S_i < M$.

In the proposed endmember selection model, λ is defined as a threshold to avoid superfluous endmember and accelerate convergence rate of classification. Generally, $0 < \lambda \ll 1$. A larger λ will lead to endmember miss-selection, while a smaller one will result in the redundancy of endmember set. In this paper, we choose $\lambda = 0.01$.

Definition of endmember set E is not a trivial task. For real hyperspectral image, land cover type is too complex to be represented by a single standard pure spectral signature. In this paper, average spectral of pixels belonging to class i will be taken as the endmember spectral of this class. The average spectral is given in Eq.(3).

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_j^{(i)} \quad (3)$$

where N_i is the number of pixels in class i .

Spectral unmixing based on obtained B_i will provide soft labels for original labeled samples. Linear spectral unmixing is adopted since it is considered without significant loss of information in the majority of applications.

As a typical linear unmixing model, fully constrained least squares (FCLS) unmixing has been widely used to achieve fractional abundance of pixels. Due to its advantage of satisfying abundance constraints and least square error, FCLS is utilized for spectral unmixing in the proposed method.

2.2 Soft MLR-Based Classification

A soft classification technique, based on multinomial logistic regression (MLR), is exploited to take advantages of the provided soft labels (fractional abundance) to offer more reliable probabilistic output.

Multinomial logistic regression (MLR) models posterior probabilistic as Eq.(4)

$$P(y_i^t = 1 | x_i, \omega) = \frac{\exp(\omega^{(i)T} x_i)}{\sum_{j=1}^M \exp(\omega^{(j)T} x_j)} \quad (4)$$

where $X = [x_1, x_2, \dots, x_N]$, $x_i \in \mathbb{R}^D$ is a D -dimensional pixel in hyperspectral image. $Y = [y_1, y_2, \dots, y_N]$ and $y_i = [y_i^1, y_i^2, \dots, y_i^M] \in \mathbb{R}^M$ describes the label of x_i . If only one label exists in y_i (e.g. $y_i = [0, 1, 0, 0]$), x_i is a pure pixel (e.g. x_i belongs to class 2). Otherwise, x_i is a mixed pixel composed of several classes. M is the number of class, N is the number of pixels. $\omega = [\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(M)}]^T \in \mathbb{R}^{M \times D}$ denotes regressors.

A small set of labeled samples $\{(y_1, x_1), (y_2, x_2), \dots, (y_L, x_L)\}$ is used to estimate logistic regressors ω through maximum a posteriori (MAP) estimation in Eq.(5) and Eq.(6).

$$\hat{\omega} = \arg \max_{\omega} \{l(\omega) + \log p(\omega)\} \quad (5)$$

$$l(\omega) = \log \prod_{i=1}^L p(y_i | x_i, \omega) = \sum_{i=1}^L \left(\sum_{t=1}^M y_i^{(t)} \omega^{(k)T} x_i - \log \sum_{t=1}^M \exp(\omega^{(k)T} x_i) \right) \quad (6)$$

where $l(\omega)$ acts as a log-likelihood function of labeled samples and $p(\omega)$ is the prior of ω .

Since the labeled sample set is quite small and unlabeled samples are not adopted in this model, prior of ω is hardly to be estimated to fit training samples. Therefore, in this paper, $p(\omega)$ is not taken into account in MAP estimation of ω .

A soft sparse MLR model proposed in [9] is capable of dealing with soft label problem, where the regressors can be efficiently learnt by the LORSAL [10]. In the proposed method, estimated fractional abundance of labeled sample is regarded as the soft labels for soft sparse MLR classifier.

2.3 Outline of the Proposed Method

The proposed classification method, which jointly uses spectral unmixing, is an iterative learning process shown as Table 1. By updating endmember set of each class and estimating soft labels of training samples, soft MLR-based classifier is optimized to improve classification accuracy after each cycle.

For the proposed method, hard classification method, such as maximum likelihood classifier (MLC) can be exploited to initialize confusion matrix. The iteration process will stop until the iteration number is satisfied or the increasing of classification accuracy is weakened.

Table 1. Outline of the proposed method. OA is overall accuracy and AA is average accuracy

Algorithm: Adaptive endmember selection for classification

Inputs: C :confusion matrix; E :endmember set of hyperspectral image

Outputs: ω :optimal regressors; classification results (OA, kappa, AA)

Initialization: determine C by reliable traditional classification (*e.g.* MLC)

Repeat:

- (1) Select endmember set per class B_i via Eq.(1), Eq.(2);
- (2) Generate fractional abundance according to FCLS unmixing for training sample
- (3) Optimize regressors ω with Eq.(6) and classify hyperseptal image.

Until iterated condition is satisfied

3 Experimented Results and Analysis

In this paper, experiments are performed on one synthetic data set and two real data sets: AVIRIS Indian Pine data and ROSIS Pavia University data. Superiority of the proposed method is demonstrated by comparing with probabilistic SVM and spectral information divergence (SID). Principle component analysis (PCA) is adopted to reduce the dimensionality of features.

3.1 Experiments on Simulated Data

A synthetic data set is generated by pure spectral signatures, which is composed of 420 bands with wavelength ranging from $0.4\mu\text{m}$ to $2.5\mu\text{m}$, selected from U.S. Geological Survey (USGS) digital spectral library [11]. The spatial information is provided by the GIS map of AVIRIS Indian Pine data. With 10 classes, total number of pixels in simulated data is 8311. $N(0, I)$ Gaussian noise is added to the scene to get a SNR of 40dB. For all pixels of class i , 10% pixels are randomly generated as pure ones, 90% pixels are mixed by fixed hybrid mode according to linear mixture model. The fixed hybrid mode (*e.g.* class 1 is mixed with class 5) is designed to simulate the mixed mode in real hyperspectral data. For every mixed pixel, we assign 50%~60% abundance to the endmember of predominate class, remaining abundance is randomly assigned to endmembers of the mixed classes.

The threshold of PCA is 0.99. For classes which have less than 100 labeled pixels, we randomly choose 5 labeled samples as training ones. Otherwise 1% labeled pixels per class is training samples. 20 iterations are operated for every Monte Carlo run.

Table 2(left three columns) indicates that, for one random Monte Carlo run, the proposed method, which has an overall accuracy of 88.14%, can achieve better results than traditional supervised methods. Effectiveness can also be concluded in kappa and average accuracy. For 10-independent Monte Carlo runs, the proposed method has increased by 4.41% in overall accuracy (OA) and 13.59% in average accuracy (AA) compared to probabilistic SVM. Additionally, the proposed method is 4.05% higher in OA and 2.87% higher in AA than SID (see Table 3 left three columns).

3.2 Experiments on Real Data

Two real hyperspectral images are used in our experiments.

- **AVIRIS Indian Pine data.** This data represents the agricultural area of Indian Pine in the northern part of Indian. Composed of 224 bands (ranging from 0.4-2.5 μ m), this image has 10nm spectral resolution and 20m spatial resolution [12]. It contains 16 mutually exclusive classes and 145· 145 pixels, in which a total of 10366 pixels are labeled. This dataset is challenging for classification due to the presence of mixed pixels in all available classes and unbalanced number of labeled pixels per class [4].
- **ROSIS Pavia University data.** This data describes urban area of the University of Pavia, Italy. It has total 103bands (ranging from 0.43-0.86 μ m) and 1.3m spatial resolution. It contains 9 mutually exclusive classes and 610· 340 pixels, in which 42776 pixels are labeled. This dataset help to validate the proposed method.

Threshold of PCA is 0.999. For AVIRIS data, training samples selection strategy is same as the simulated data. 0.5% labeled pixels per class will form training samples for ROSIS data. Also, 20 iterations are operated for every Monte Carlo run.

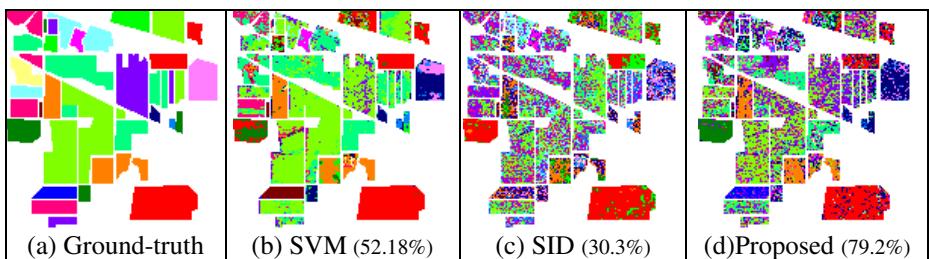


Fig. 2. Classification maps and OA results of the proposed method and contrasting approaches for a random Monte Carlo run on AVIRIS Indian Pine data set

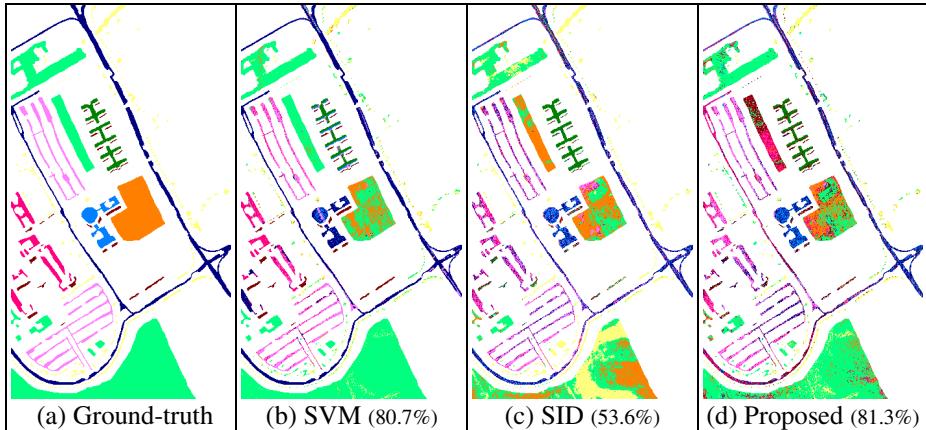


Fig. 3. Classification maps and OA results of the proposed method and contrasting approaches for a random Monte Carlo run on ROSIS Pavia University data set

Fig. 2 shows classification maps of AVIRIS Indian Pine data obtained by different classification methods for one random Monte Carlo run. Ground-truth map of AVIRIS data can be seen in Fig. 2(a). Fig. 3 is classification maps of ROSIS Pavia University data. Fig. 3(a) represents the ground-truth map.

Table 2 shows the classification accuracy of each class for one random Monte Carlo run on two real hyperspectral data sets. The proposed method has separately increased by 11.92% and 38.8% in OA for AVIRIS data compared with probabilistic SVM and SID based methods (see middle three columns). Similarly, the superiority of proposed method is obvious in kappa and AA. For ROSIS data, the proposed method is 14.45% higher in AA than the probabilistic SVM.

Table 2. Classification result (%) of three data sets compared with probabilistic SVM and SID

	Simulated data			AVIRIS Indian Pine			ROSIS Pavia University		
	SVM	SID	Proposed	SVM	SID	Proposed	SVM	SID	Proposed
Class 1	100.0	76.71	100.0	88.89	46.30	98.15	88.48	51.67	71.42
Class 2	100.0	100.0	100.0	67.71	21.83	95.26	97.06	46.92	85.89
Class 3	9.830	100.0	39.32	37.29	25.90	25.50	66.03	35.64	100.0
Class 4	92.56	44.67	99.40	1.710	13.25	26.92	72.49	96.05	74.09
Class 5	77.64	90.09	70.01	38.23	1.610	60.97	69.07	97.55	77.10
Class 6	92.31	92.31	92.31	80.19	29.99	64.26	44.02	53.95	55.76
Class 7	15.13	90.18	67.69	88.46	61.54	100.0	20.45	46.39	96.09
Class 8	90.08	69.11	89.98	26.18	32.31	34.15	70.75	59.75	98.59
Class 9	100.0	76.26	91.00	19.01	60.00	40.00	97.99	24.08	97.47
Class 10	82.25	100.0	77.52	80.06	13.43	94.42	-	-	-
Class 11	-	-	-	81.77	27.15	91.25	-	-	-
Class 12	-	-	-	9.930	25.08	17.43	-	-	-
Class 13	-	-	-	1.892	28.77	70.75	-	-	-
Class 14	-	-	-	97.53	75.04	74.11	-	-	-
Class 15	-	-	-	16.58	25.00	19.74	-	-	-
Class 16	-	-	-	87.37	60.00	72.63	-	-	-
OA	87.52	80.53	88.14	57.18	30.30	69.10	80.70	53.62	81.34
kappa	87.32	80.26	87.94	56.81	29.87	68.88	80.20	52.73	80.94
AA	75.98	83.93	82.72	52.56	34.30	61.28	69.59	56.89	84.04

Table 3. Average classification accuracy (%) of 10 independent Monte Carlo runs obtained by the proposed method and contrasting ones on three data sets

	Simulated data			AVIRIS Indian Pine			ROSIS Pavia University		
	SVM	SID	Proposed	SVM	SID	Proposed	SVM	SID	Proposed
OA	82.84	83.20	87.25	56.21	27.64	67.73	80.55	51.52	79.15
kappa	82.56	82.95	87.03	55.81	30.15	67.53	80.04	50.61	78.71
AA	68.39	79.17	81.98	50.86	27.24	62.19	67.44	56.36	82.38

Table 3 statistics average classification result (OA, kappa, AA) for 10-independent Monte Carlo runs on two real hyperspectral data sets (middle and right three columns). The proposed method is 11.52% higher in OA than probabilistic SVM for AVIRIS data. For ROSIS data, it is 14.94% higher in AA than that of the probabilistic SVM.

4 Conclusion

In this paper, a novel method, which jointly uses soft classification and spectral unmixing technique, is proposed for hyperspectral image classification. Confusion matrix is utilized for endmember selection. Then obtained endmember set is adopted for FCLS unmixing. FCLS unmixing estimated soft labels are exploited to optimize soft MLR classifier. This procedure is executed iteratively to get required performance. Experimental results on synthetic and real hyperspectral data sets show the effectiveness of the proposed method compared with probabilistic SVM and other traditional methods. Future work will be devoted to investigate advanced methods to get representative endmember set of the image.

Acknowledgments. This work is supported by the State Key Program of National Natural Science of China (No.61231016), National Natural Science Foundation of China (No.61272288, No.61201291), NPU Foundation for Fundamental Research (No.JCT20130108).

References

1. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J.: A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing* 5(3), 606–617 (2011)
2. Plaza, A., Martin, G., Plaza, J., Zortea, M., Sanchez, S.: Recent Developments in Spectral Unmixing and Endmember Extraction. In: *Optical Remote Sensing-Advances in Signal Processing and Exploitation Techniques*. Springer, New York (2010)
3. Villa, A., Chanussot, J., Benediktsson, J.A., Jutten, C.: Spectral Unmixing for the Classification of Hyperspectral Images at a Finer Spatial Resolution. *IEEE Journal of Selected Topic in Signal Processing* 5(3), 521–533 (2011)
4. Villa, A., Li, J., Plaza, A., Bioucas-Dias, J.M.: A New Semi-supervised Algorithm for Hyperspectral Image Classification based on Spectral Unmixing Concepts. In: *2011 3rd Workshop on Selected Topics in Signal Processing, Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS* (2011)

5. Plaza, A., Martinez, P., Perez, R., Plaza, J.: A New Approach to Mixed Pixel Classification of Hyperspectral Imagery based on Extended Morphological Profiles. *Pattern Recognition* 37, 1097–1116 (2004)
6. Böhning, D.: Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics* 44, 197–200 (1992)
7. Li, J., Bioucas-Dias, J., Plaza, A.: Semi-supervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression with Active Learning. *IEEE Transactions on Geoscience and Remote Sensing* 48, 4085–4098 (2010)
8. Heinz, D., Chang, C.-I.: Fully Constrained Least Squares Linear Mixture Analysis for Material Quantification in Hyperspectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 39, 529–545 (2001)
9. Li, J., Bioucas-Dias, J., Plaza, A.: Semi-supervised Hyperspectral Classification Using Soft Labels. In: *IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS 2011* (2011)
10. Krishnapuram, B., Carin, L., Figueiredo, M., Hartemink, A.: Sparse Multinomial Logistic Regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(6), 957–968 (2005)
11. USGS digital spectral library online, <http://speclab.cr.usgs.gov>
12. Meng, Q.J., Zhang, Y.N., Wei, W., Ren, Y.M., She, H.W.: Class-specific Artificial Immune Recognition Method for Hyperspectral Image Classification. In: *2012 IEEE 11th International Conference on Signal Processing (ICSP)*, pp. 851–855 (2012)

A Subarea-Location Joint Spelling Paradigm for the BCI Control

Erwei Yin, Jun Jiang, Yang Yu, Jingsheng Tang, Zongtan Zhou, and Dewen Hu

College of Mechatronic Engineering and Automation,
National University of Defense Technology, Changsha,
Hunan 410073, P.R. China
dwhu@nudt.edu.cn

Abstract. Brain computer interface (BCI) speller is an important issue in BCI research. In this paper, we propose a novel spelling paradigm for enhancing the performance of BCI speller. In our approach, the target character is detected by the combination of the P300 potential and the steady-state visual evoked potential (SSVEP). Specifically, the P300 detection mechanism and the SSVEP detection mechanism are employed as two sub-spellers for identifying the number of the subarea and location of target character, respectively and simultaneously. The experimental results show that the information transfer rate (ITR) of our BCI system was significantly improved compared to the traditional BCI approaches, i.e. P300 speller and SSVEP speller.

Keywords: brain computer interface (BCI) speller, hybrid BCI, electroencephalogram (EEG), P300, steady-state visual evoked potential (SSVEP).

1 Introduction

Brain-computer interface (BCI) is a kind of system that can directly acquire signals from the human brain and translates them into digital commands which can be recognized and processed on a computer. It provides a reestablishing communication pathway and environmental control capability to severely disabled persons [1]. Up to now, many researchers have focused on BCI spellers having an alphabetic writing system [2]. Several different EEG signals can be used for BCI control, such as P300 event-related potential, steady-state visual evoked potential (SSVEP) and sensorimotor rhythm (SMR) [3]. However, the SMR-based BCI speller is inaccurate and requires extensive training. P300 and SSVEP potentials have been widely used in BCI speller due to the superior performance. Here, the P300 potential is elicited by rare, task-relevant events and is often recorded at approximately 300 ms after the presented stimuli in an 'oddball' paradigm [4], as well as the SSVEP potential is a periodic response evoked by a visual repetitive stimulus repeating in the range 6-60 Hz [5].

To enhance the system performance, several P300 optimal approaches were mainly focus on the stimulus presentation paradigm design, which was used to

enhance the spelling speed by decreasing the number of flashes per trial [6–8]. These approaches indeed improved the spelling speed, while this improvement also impaired the classification accuracy per trial, because the decrease of P300 amplitude [9]. In the SSVEP studies, most approaches were to apply improved signal processing and the optimal design of stimulus frequencies, which have induce significant improvement of the information transfer rate (ITR) [10–12]. Unfortunately, the classification accuracy tends to decrease by the increasing of the number of options, as well as the number of options in SSVEP-based speller are limited by the number of stimulus frequencies in the case of use the PC monitor [11], which restricts the further enhancement of the ITR. So far, most studies of BCI spellers are explored only rely on the single model brain signal, which is facing the bottleneck of the ITR enhancement due to the tradeoff between speed, accuracy and number of options.

Recent works have validated a new framework for BCI research called the hybrid BCI. A hybrid BCI is a system that combines more than one different BCI approach [13]. Allison *et al* developed a hybrid BCI based on the SMR and the SSVEP [14]. Using this approach, the number of illiterate subjects was reduced by the improvement of the accuracy and reducing the selection time. The combination of SMR and P300 signals for target selection in 2-D cursor control was presented by Li *et al* [15]. The target selection was more accurate compared with that resulting from either the SMR or P300 feature alone. Moreover, a hybrid BCI speller based on the information fusion of SSVEP and P300 was proposed in our previous work [16]. The BCI speller could achieve a higher accuracy compared with the conventional BCI speller. These positive results of combined approaches have motivated us to study the hybrid BCI speller.

In this paper, we propose a subarea-location joint (SLJ) spelling paradigm which is a kind of hybrid BCI approach. The target character here is detected by the combination of P300 and SSVEP potentials. More specifically, the P300 detection mechanism and the SSVEP detection mechanism are employed as two sub-spellers for identifying the number of subarea and location of target character respectively and simultaneously.

2 Materials and Methods

2.1 Stimulation Paradigm

The target character of our paradigm is detected by 2-dimensional time-frequency features, including the P300 feature and the SSVEP feature respectively. To evoke these two brain potentials simultaneously, the flash pattern mechanisms presented here are composed of random flashings and periodic flickers. The random flashings are devised by highlighting the items using orange crosses in a pseudorandom sequence, as well as the periodic flickers are created using the white rectangular objects whose appearance and disappearance alternated on a black background. The additional details of these hybrid stimuli mechanisms were described in our previous study [16].

As can be seen from figure 1, we employ classical standard of the BCI speller with 6×6 matrix. The character matrix is divided to six subareas surrounded by the yellow broken lines. All the items of each subarea flash at same frequency. The six frequencies are set at 8.18, 8.97, 9.98, 11.23, 12.85 and 14.99 Hz respectively. The selection of these frequencies was discussed thoroughly in our previous study. The SSVEP feature is used to identify the subarea the target item belongs to. At the same time, the orange crosses are highlighted at the same location of all the subareas. The location of the target item in the subareas is determined by the P300 feature. Thus, the target item is detected a 2-dimensional coordinate composed by the number of the subarea and its location.

In this stimulation paradigm, we only employ 6-flash pattern of the P300 and 6 frequencies of the SSVEP to achieve the spelling of each of 36 items. Thus the stimulus time decreases to half of those of conventional P300 speller, as well as the selections increases to six times of the SSVEP speller.

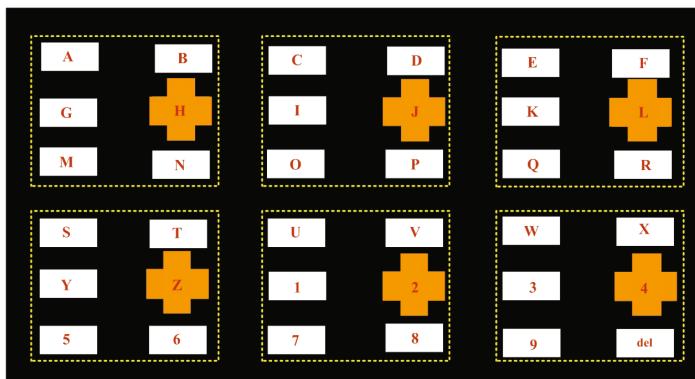


Fig. 1. Illustrations of the stimulus configuration

2.2 Experimental Setup

Subjects. Six healthy subjects (2 females and 4 males, age 18-32 years, with a mean age of 27.2 years) participated in the study. All of the subjects had normal or corrected-to-normal vision. Only one of the subjects had used the P300-SSVEP based hybrid BCI prior to the study, and the remaining five subjects were complete novices. All subjects signed an informed consent form in accordance with the Declaration of Helsinki.

Data Collection. The EEG signals were recorded using a BrainAmp DC Amplifier (Brain Products GmbH, Germany). Using the 64-channel extended international 10/20 system, ten-channel active electrodes were placed at Fz, Cz, Pz, P3, P4, Oz, O1, O2, PO7, and PO8, referenced to P8 and grounded to Fpz. For P300 detection, only the Fz, Cz, Pz, P3, P4 and Oz channels were considered, and the Oz, O1, O2, POz, PO7 and PO8 channels were selected for the SSVEP

detection. Each of the impedances was kept below $10\text{ k}\Omega$ prior to recording. The EEG signals were sampled at 250 Hz and filtered using a 50-Hz notch filter. The data collection and experimental procedure associated with the BCI speller were controlled using the BCI2000 platform [17], which provides a Python interface for stimuli presentation and a MatLab interface for signal processing.

Experimental Paradigm. The experiments were performed in a normal office room. The subjects were seated in front of LED monitor with a refresh rate of 60 Hz. Each run was consisted by 14 characters' spellings, as well as eight trials considered for the detection of one character. The stimulus onset asynchrony (SOA) of the random flashes was 240 ms, which means that each flash was highlighted for 120 ms, with a 120-ms delay between flashes. Here one trial was defined as a complete cycle of flashes in which all of stimulus code performed once. Furthermore, the orders of the sessions were picked up in a random order for each subject to avoid fatigue bias, as well as a 5-min break was given for rest after each run. To avoid confusing the subjects, a 2-s break was given to allow the subjects to locate the next symbol.

In the P300 session, we employed the standard P300 speller with RC paradigm. The random flashing was presented also using the orange crosses. Here one trial was consisted by 12 flashes. The subjects were instructed to maintain a mental count of the number of times the prompted symbol was highlighted. Each subject was required to perform six runs. In the SSVEP session, all of the cells of the matrix flickered at six different frequencies with the arrangement as the frequencies in our SLJ paradigm. Only three runs were performed because SSVEP detection did not need the train runs. During these runs, each subject was instructed to gaze at the prompted symbol. Each trial here had the same time length with 6 flashes. In the hybrid session, the SLJ stimulus paradigm was employed for evoking the two brain potentials, as well as the subjects were instructed to synchronously conduct the tasks performed in both P300 and SSVEP condition. Here, each subjects also performed six runs.

2.3 Signal Processing

In the SLJ spelling paradigm, we explored parallel P300 and SSVEP signal processing (see figure 2). To begin with, the EEG signals were divided into two kinds by the P300 channels and the SSVEP channels. Secondly, these two kinds of signals were processed by the P300 and SSVEP detection mechanisms respectively and simultaneously. Finally, the target coordinate was determined based on the maximum score of the time-frequency features including P300 and SSVEP. Here the coordinate defined as the number of the subarea and its position. The complete signal processing approaches are summarized in the following.

P300 Detection. First, the EEG data of P300 channels were filtered using a 0.1-45-Hz bandpass filter to eliminate the signal excursion and high-frequency noise. Next, the 800-ms epochs of the data, starting from the onset of each stimulus, were extracted for the P300 feature analysis. Because of the high sampling

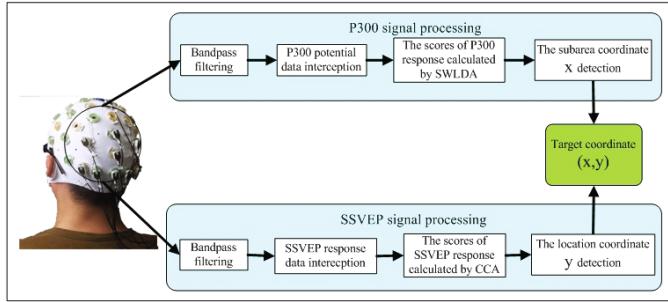


Fig. 2. The signal processing of our hybrid BCI speller

rate of the signal recording relative to the low frequency of the information received for detection of the P300 response, the data were down-sampled from 250 to 25 Hz to reduce the feature size by selecting every tenth sample from the EEG. Step-wise linear discriminant analysis (SWLDA) was performed to train a classifier to calculate the scores of the P300 responses [18]. After that, the scores of each stimulus code were calculated by averaging all of the scores according to the same code numbers. The stimulus code with the maximum average score was defined as the first dimensional coordinate of the target x .

SSVEP Detection. To begin with, the data collected from the SSVEP channels were filtered using a 4-35-Hz bandpass filter. Then, we used canonical correlation analysis (CCA) approach to calculate the correlation between the stimulus frequency and the multi-channel EEG data [19]. Here, the stimulus frequency was represented as a square-wave periodic signal which was decomposed into Fourier series of its harmonics. Finally, the stimulus frequency with the maximum correlation coefficient was regarded as the target frequency. That is, the number of the frequency was defined as the second dimensional coordinate of target y .

Thus, the target character was sited by the 2-dimensional coordinate (x, y) .

3 Experimental Results

To evaluate BCI performance, we computed both the classification accuracy and ITR. ITR is widely used in the BCI community as an important metric, which is computed by the following formula [20].

$$ITR = \left\{ \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1 - P}{N - 1} \right\} / T \quad (1)$$

where N represents the number of options, P denotes the classification accuracy, and T is the time interval per selection. Here T is computed from

$$T = (S \cdot L \cdot R + I) / 60 \quad (2)$$

where S denotes the SOA (240 ms), L is the number of flashes per trial, R is the number of trials per symbol, and I is the time break between selections (2 s).

The performance comparisons between the SLJ speller, P300 speller and SSVEP speller are shown in figure 3. The ITR of the SLJ speller was significantly better compared with both P300 and SSVEP speller (see the upper half of the chart). In particular, the maximum value of average ITR was improved by 43.24% (SLJ speller: 39.04 bits/min, at 3rd trial vs. P300 speller: 27.25 bits/min, at 1st trial) and 96.54% (SLJ speller: 39.04 bits/min, at 3rd trial vs. SSVEP speller: 19.86 bits/min, at 3rd trial) respectively. To achieve effective communication, an accuracy of over 70% is commonly required [21]. Note that, the average accuracy of P300 speller only reached to 60.32% at the 1st trial (see the lower half of the chart), which was inefficient for BCI control. While, the accuracies of SLJ and SSVEP speller achieved 88.10% and 92.06% respectively, at their maximum average performances. It means that these two spellers can provide effective control for spelling. Furthermore, the error bars in the figure represent the minimum and maximum ITR and accuracy achieved according to the number of trials. As illustrated by the error bars, the best performance of the subjects reached to 57.17 bits/min with the accuracy of 95.24%, by using the SLJ speller.

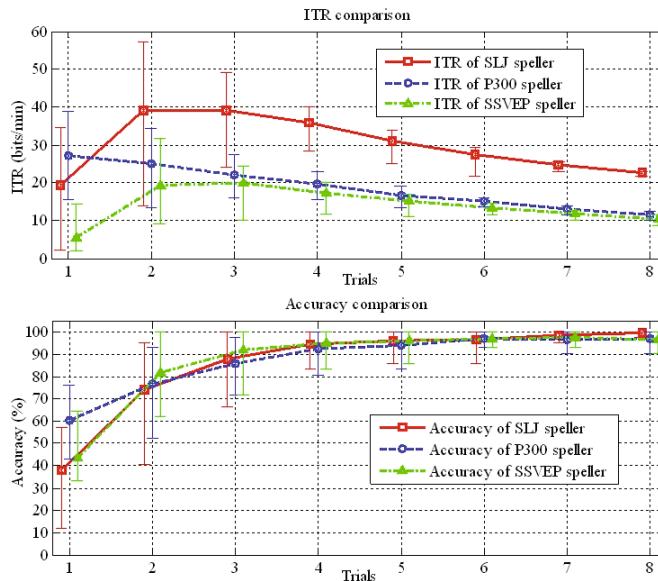


Fig. 3. The system performance comparisons between the SLJ speller, P300 speller and SSVEP speller. For clarity, the curves are shifted slightly left and right.

4 Conclusions and Future Work

In this paper, we proposed a novel hybrid BCI approach based on the SLJ spelling paradigm. In our approach, we devised a hybrid stimulus mechanism composed of random flashings and periodic flickers. Furthermore, the target character was identified based on 2-dimensional time-frequency features, including the P300 and SSVEP. The experimental results showed that the performance of our approach was significantly better than the traditional single-model BCI approaches: P300 and SSVEP speller, due to the information superposition of the two kinds of brain potentials. The maximum value of average ITR achieved to 39.04 bits/min by using the SLJ speller. Importantly, the highest peak performance measured up to 57.17 bits/min with the accuracy of 95.24%, which is worth mentioning that the SLJ speller is promising for practical application of the stimulus-driven BCI system.

In future work, we will focus on the optimal matrix size and spelling speed design based on the overall performance of each subject for our speller. When using the P300 or SSVEP potential, different subjects perform different spelling abilities. Because the target detection of our speller depends on the accuracies of P300 and SSVEP simultaneously, one of these two potentials exporting bad accuracy will directly disturb the target detection. Thereby, the optimal matrix size design for each subject will bring more adequate utilization of the multi-model information thus achieve better performance. Furthermore, the subjects' performance even varied rapidly between the symbols [22]. Thus, the study of the adaptive speller concept, in which the spelling speed may vary depending on the current user's state, is another important future direction of study.

Acknowledgments. This work was supported in part by the National High Technology Research and Development Program (Project 2012AA011601) and the National Basic Program of China (Project 2011CB707802).

References

1. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-Computer Interfaces for Communication and Control. *Clin. Neurophysiol.* 113, 767–791 (2002)
2. Mak, J.N., Arbel, Y., Minett, J.W., McCane, L.M., Yuksel, B., Ryan, D., Thompson, D., Bianchi, L., Erdogmus, D.: Optimizing the P300-Based Brain-Computer Interface: Current Status, Limitations and Future Directions. *J. Neural Eng.* 8, 025003 (2011)
3. Cecotti, H.: Spelling with Non-invasive Brain Computer Interfaces-Current and Future Trends. *Journal of Physiology-Paris* 105, 106–114 (2011)
4. Farwell, L.A., Donchin, E.: Talking off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Brain Potentials. *Clin. Neurophysiol.* 70, 510–523 (1988)
5. Vialatte, F.B., Maurice, M., Dauwels, J., Cichocki, A.: Steady-State Visually Evoked Potentials: Focus on Essential Paradigms and Future Perspectives. *Prog. Neurobiol.* 90, 418–438 (2010)

6. Jin, J., Allison, B.Z., Sellers, E.W., Brunner, C., Horki, P., Wang, X., Neuper, C.: An Adaptive P300-Based Control System. *J. Neural Eng.* 8, 036006 (2011)
7. Allison, B.Z., Pineda, J.A.: Effects of SOA and Flash Pattern Manipulations on ERPs, Performance, and Preference: Implications for a BCI System. *Int. J. Psychophysiol.* 59, 127–140 (2006)
8. Townsend, G., LaPallo, B.K., Boulay, C.B., Krusienski, D.J., Frye, G.E., Hauser, C.K., Schwartz, N.E., Vaughan, T.M., Wolpaw, J.R., Sellers, E.W.: A Novel P300-Based Brain-Computer Interface Stimulus Presentation Paradigm: Moving Beyond Rows and Columns. *Clin. Neurophysiol.* 121, 1109–1120 (2010)
9. Gonsalvez, C.L., Polich, J.: P300 Amplitude is Determined by Target-to-Target Interval. *Psychophysiology* 39, 388–396 (2002)
10. Gao, X., Xu, D., Cheng, M., Gao, S.: A BCI-Based Environmental Controller for the Motion-Disabled. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11, 137–140 (2003)
11. Bin, G., Gao, X., Yan, Z., Hong, B., Gao, S.: An Online Multi-Channel SSVEP-Based Brain-Computer Interface Using a Canonical Correlation Analysis Method. *J. Neural Eng.* 6, 46002 (2009)
12. Hwang, H., Lim, J., Jung, Y., Choi, H., Lee, S., Im, C.: Development of an SSVEP-Based BCI Spelling System Adopting a QWERTY-Style LED Keyboard. *J. Neurosci. Methods* 208, 59–65 (2012)
13. Allison, B.Z., Brunner, C., Kaiser, V., Muller-Putz, G.R., Neuper, C., Pfurtscheller, G.: Toward a Hybrid Brain-Computer Interface Based on Imagined Movement and Visual Attention. *J. Neural Eng.* 7, 026007 (2010)
14. Brunner, C., Allison, B.Z., Altstatter, C., Neuper, C.: A Comparison of Three Brain-Computer Interfaces Based on Event-Related Desynchronization, Steady State Visual Evoked Potentials, or a Hybrid Approach Using both Signals. *J. Neural Eng.* 8, 025010 (2011)
15. Long, J.Y., Li, Y.Q., Yu, T.Y., Gu, Z.H.: Target Selection with Hybrid Feature for BCI-Based 2-D Cursor Control. *IEEE Trans. Biomed. Eng.* 59, 132–140 (2012)
16. Yin, E., Zhou, Z., Jiang, J., Chen, F., Liu, Y., Hu, D.: A Novel Hybrid BCI Speller Based on the Incorporation of SSVEP into the P300 Paradigm. *J. Neural Eng.* 10, 026012 (2013)
17. Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R.: BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Trans. Biomed. Eng.* 51, 1034–1043 (2004)
18. Krusienski, D.J., Sellers, E.W., Cabestaing, F., Bayoudh, S., McFarland, D.J., Vaughan, T.M., Wolpaw, J.R.: A Comparison of Classification Techniques for the P300 Speller. *J. Neural Eng.* 3, 299–305 (2006)
19. Lin, Z., Zhang, C., Wu, W., Gao, X.: Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-Based BCIs. *IEEE Trans. Biomed. Eng.* 53, 2610–2614 (2006)
20. Wolpaw, J.R., Birbaumer, N., Heetderks, W.J., McFarland, D.J., Peckham, P.H., Schalk, G., Donchin, E., Quatrano, L.A., Robinson, C.J., Vaughan, T.M.: Brain-Computer Interface Technology: A Review of the First International Meeting. *IEEE Trans. Rehabil. Eng.* 8, 161–173 (2000)
21. Pires, G., Nunes, U., Castelo-Branco, M.: Comparison of A Row-Column Speller Vs. A Novel Lateral Single-Character Speller: Assessment of BCI for Severe Motor Disabled Patients. *Clin. Neurophysiol.* 123, 1168–1181 (2012)
22. Lenhardt, A., Kaper, M., Ritter, H.J.: An Adaptive P300-Based Online Brain-Computer Interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* 16, 121–130 (2008)

A Robust Real-Time Tracking Method of Fast Video Object Based on Gaussian Kernel and Random Projection

Yajuan Feng¹, Lina Wang², and Shiyin Qin¹

¹ School of Automation Science and Electrical Engineering, Beihang University, 100191, Beijing, China

sahala337@163.com, qsy@buaa.edu.cn

² National Key Laboratory of Science and Technology on Aerospace Intelligent Control, Beijing Aerospace Automatic Control Institute, 100854, Beijing, China
violina@126.com

Abstract. It is a challenging topic how to achieve the real-time tracking of fast video object under complex environment. In this paper, a scheme and its corresponding implementing algorithms of real-time tracking of fast video object are designed and perfected, which are characterized with high performances of real-time tracking and robustness. At first, a kind of scheme is designed for the real-time tracking of fast video object and corresponding implementing strategies for some key modules are proposed. Then the particle filter is employed to predict the pose state of fast video object and the motion object is discriminated from its background by Gaussian kernel and random projection. Moreover, an adaptive feature selection method is used to enhance the robustness and tracking efficiency. A series of experiment results demonstrate that the scheme and algorithms proposed in this paper outperform the current existing algorithms in the tracking efficiency, accuracy and robustness.

Keywords: object tracking, Gaussian kernel, random projection, adaptive feature selection, real-time performance.

1 Introduction

Real-time object tracking plays more and more important role in many practical application fields such as traffic monitoring, surveillance and video indexing, etc. Over the years, a series of tracking methods have been proposed, but it is still a challenging tough task to tracking fast moving object due to the presence of partial occlusion, background clutter and dynamic changes under complex environment.

The purpose of tracking is to track the state changes of specified object in a sequence of images. To achieve this goal, two kinds of methods have developed over these years. For generative methods [1-3], tracking is formulated as a searching problem. The region which has most similarity with reference model will be considered as the best possible location of tracked object in each frame. Discriminative methods [4-6] treat object tracking as a binary classification problem, and update the trained classifier online during the tracking procedure.

In this paper, we combine generative and discriminative methods. Here, particle filter is used to predict the pose state of object. To characterize the object more efficiently, we combine random projection used in [6] with Gaussian kernel function to construct the appearance model. According to the fact that the discrimination ability of tracker is relate to the dimension of feature space, we develop a novel feature selection strategy which adaptively regulates feature space based on the discrimination between object and background. Therefore, the proposed tracking method is robust and can cope with fast moving object with relatively high frame rate.

2 Scheme and Implementing Algorithm

2.1 Scheme and Strategy of Real-Time Tracking of Fast Video Object

As is shown in Fig. 1, there are 2 feedback loops with 4 function modules in our scheme. For an image reading from video stream, the feature extraction is carried out with random projection according to the prediction of object region and dimension regulation. Driven by the adaptively selective feature vector information the object tracking is actualized with Bayesian classifiers which can effectively discriminate the object from its background in a maximum confidence degree. According to the confidence distribution, the dimension of feature space is regulated to enhance the degree of discrimination between object and its background. Meanwhile the prediction of next candidate object regions is carried out by using particle filer based on the confidence degree of current localization of object.

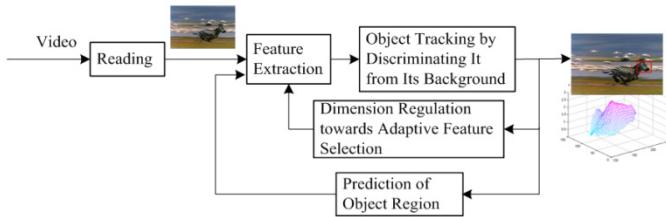


Fig. 1. The scheme of proposed tracking method

2.2 Prediction of Pose State of Motion Object Based on Particle Filter

Let $\chi_t = (x, y, H_x, H_y)$ represent the object state in t frame, where (x, y) are the center location of the tracked object and (H_x, H_y) are the corresponding size. Given all available observation variables $z_{1:t} = \{z_1, z_2, \dots, z_t\}$, particle filter approximates the posterior distribution $p(\chi_t | z_{1:t})$ by using discrete random measure which is defined by a weighted particle set $S = \{(\chi_t^i, \omega_t^i)\}_{i=1}^N$. Therefore posterior probability density can be computed by

$$S = \left\{ (\chi_t^i, \omega_t^i) \right\}_{i=1}^N p(\chi_t | z_{1:t}) = \sum_{i=1}^N \omega_t^i \delta(\chi_t - \chi_t^i). \quad (1)$$

where $\delta(\cdot)$ is the Dirac delta function and the weight of particle is updated as

$$\omega_t^i = \omega_{t-1}^i \frac{p(z_t | \chi_t^i) p(\chi_t^i | \chi_{t-1})}{q(\chi_t | \chi_{1:t-1}, z_{1:t})}. \quad (2)$$

where $p(z_t | \chi_t)$ is the observation model and $q(\chi_t | \chi_{1:t-1}, z_{1:t})$ is an important distribution which is used to generate candidate particles.

In this paper, the state variables are considered as independent with each other. The transition model $p(\chi_t | \chi_{t-1})$ is constrained by assuming a Gaussian distribution. The observation model $p(z_t | \chi_t)$ reflects the possibility of candidate region becoming object, which is computed by classifier. After classification, the importance weight of each particle is updated and normalized according to Equation 2. Particles are then resampled according to their respectively weights to avoid degeneracy.

2.3 Discrimination of Motion Object from Its Background with Gaussian Kernel and Random Projection

Random projection is an efficient dimensionality reduction technique in which the high dimensional data is projected to a much lower dimensional space without losing information. For a candidate image region I_r , a multi-scale filter bank is defined as

$$H = \left\{ h_{w,h}(x, y) \right\}_{0 < w < H_x, 0 < h < H_y} \quad (3)$$

where $h_{w,h}(x, y)$ is a rectangle filter with following formula

$$h_{w,h}(x, y) = \begin{cases} 1 & 0 \leq x \leq w \text{ and } 0 \leq y \leq h \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

(H_x, H_y) and (w, h) are the size of image region and rectangle filter respectively. The high dimensional representation of image region is given as $V = (v_1 v_2 \cdots v_d)^T \in \mathbf{R}^d$ where $v_i = \sum_{(x,y) \in \Psi(x,y)} I(x, y)$ and $\Psi(x, y) = \{(x, y) | h_{w,h}(x, y) = 1\}$. Using random projection [7]

$$F = RV. \quad (5)$$

where \mathbf{R} is a random projective matrix with size of $k \cdot d$ and its element r_{ij} is assigned to conform the Gaussian distribution, thus the k-dimensional ($k \ll d$) feature vector \mathbf{F} is obtained, which represents the candidate region in compressive domain [6]. Theoretically, if the dimension of vector \mathbf{V} is high enough, which is inappropriate for real-time tracking, \mathbf{F} can characterize image region uniquely [8].

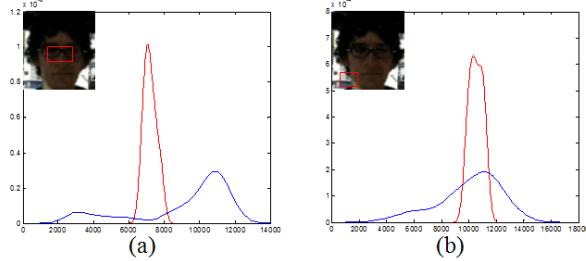


Fig. 2. Probability distributions corresponding to two different rectangle filters. Red and blue lines denote the distributions estimated based on the positive samples and negative samples, respectively. Therefore, Fig. 2(a) shows a higher degree of discrimination between object and its background.

Normally, the peripheral pixels of an image patch are the least reliable, being often affected by occlusions or disturbances from background [9]. As shown in Fig. 2, for a rectangle filter distributed in image region randomly, its distance from center affects the reliability as well. From this point, the Gaussian kernel function can be used to improve the effectiveness of current feature vector.

The distance of a rectangle filter is defined by

$$\mathbf{d}_i = \left(\frac{\sum_j \text{sign}(r_{ij})^2 x_j}{\sum_j \text{sign}(r_{ij})^2}, \frac{\sum_j \text{sign}(r_{ij})^2 y_j}{\sum_j \text{sign}(r_{ij})^2} \right). \quad (6)$$

where (x_j, y_j) are coordinate center of rectangle filter. Therefore an appearance model represented by $\mathbf{F} = \{(f_i, \mathbf{d}_i)\}_{i=1}^k$ is obtained after feature selection. We treat each element f_i in \mathbf{F} as a weak classifier h_k with four parameters $(\alpha_1, \sigma_1, \alpha_0, \sigma_0)$, and combine them as a strong classifier $H(\mathbf{F})$

$$H(\mathbf{F}) = \sum_{i=1}^k \left(k(\|\mathbf{d}_i - \mathbf{d}_0\|) \right) \log \left(\frac{p(f_i | y=1)}{p(f_i | y=0)} \right). \quad (7)$$

where $k(\|\mathbf{d}_i - \mathbf{d}_0\|) = \exp \left\{ -\frac{\|\mathbf{d}_i - \mathbf{d}_0\|^2}{(2\sigma)^2} \right\}$ is the Gaussian kernel function, $y \in \{0, 1\}$ and \mathbf{d}_0 is center of kernel function. In this paper, \mathbf{d}_0 means the coordinate center of image

region and σ is bandwidth. We assume that $p(f_i | y=1) \sim N(\alpha_1^i, \sigma_1^i)$ and similarly for $y=0$. When new data $\{(F^1, y_1), (F^2, y_2), \dots, (F^N, y_N)\}$ arrive, the following update rules will be used

$$\alpha_1^i = \gamma \alpha_1^i + (1-\gamma) \frac{1}{N_p} \sum_{n|y_n=1} f_i^n . \quad (8)$$

$$\sigma_1^i = \gamma \sigma_1^i + (1-\gamma) \sqrt{\frac{1}{N_p} \sum_{n|y_n=1} (f_i^n - \alpha_1^i)^2} . \quad (9)$$

where N_p is the number of positive samples. γ is a learning rate parameter and the update rules for α_0 and σ_0 are similarly defined. After classification, each candidate sample is assigned a confidence value which is entered in a confidence map. The position of object is then obtained through analyzing this confidence map.

2.4 Robust Tracking of Fast Object Using Adaptive Feature Selection

For robust tracking, tracker needs to on the one hand handle all possible appearance changes of object and cope with severe disturbances from background on the other, the fixed object representation always results in drifting and finally tracking failure [10], so it is necessary for tracker to be adaptive.

In the top row of Fig. 3(a), the dimension of F is 30 and there exists many interference points on the confidence map, which will weaken the discrimination capability of tracker and result in failure. However in top row of Fig. 3(b), when k is set to 70, the confidence map is more straightforward, which makes the construction of classification hyper-plane is easier for classifier. But the higher dimension of feature space is time-consuming and not necessary when the environment around object is relatively clean as shown in the bottom row of Fig. 3. Therefore, we adopt such a strategy which adaptively regulates the dimension of feature space according to the environment around object.

Based on experiments, the standard deviation is used to characterize the dispersion degree D_f^t of a confidence map at time t [11]. D_f^t is defined as follows

$$D_f^t = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\omega_t^i - \frac{1}{N} \right)^2} . \quad (10)$$

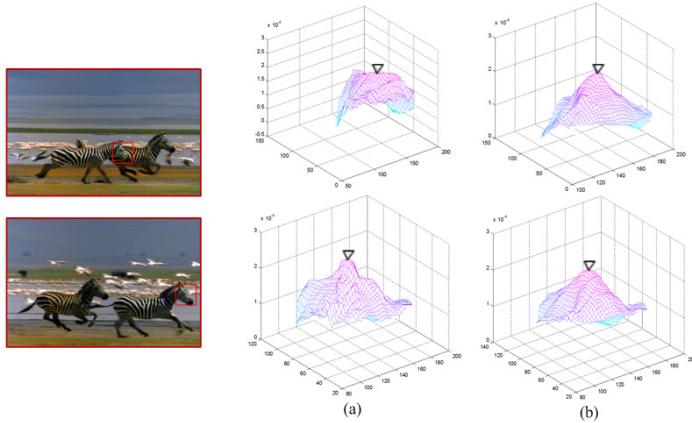


Fig. 3. Confidence maps under different environments. Classifiers in the same row are trained by same training samples with different dimensions k of \mathbf{F} : (a) $k = 30$; (b) $k = 70$.

The higher D_f means more interference points which result in drifting and tracking failure. Dimension of feature space will then be regulated by

$$k_{t+1} = k_t + m \log \left(\frac{D_f^t}{D_f^{t-1}} \right). \quad (11)$$

where m is step width and $\log(D_f^t / D_f^{t-1})$ measures efficiency of current feature space with dimension of k_t . $\log(D_f^t / D_f^{t-1})$ is positive when there are severe disturbances, therefore the dimension of feature space in next frame will be increased.

3 Experiment Results and Comparative Analysis

3.1 Experiment Data

The proposed tracking algorithm is evaluated by using four sequences which contains many challenges, such as partial occlusion(*Tiger* [12], *David Indoor* [12]), background clutter(*Tiger*, *David Indoor*, *Bicycle* [13], *Zebra* [13]), dynamic changes in background(*David Indoor Bicycle*, *Zebra*). The method is compared with four latest state-of-art tracking methods named Online-AdaBoost(OAB) [14], Compressive Tracking(CT) [6], SemiBoost(SemiB) [14] and fragment tracker(Frag) [15].

We set particle number $N = 300$ in experiment. The learning rate $\gamma = 0.85$ and the bandwidth of Gaussian kernel function $\sigma = 1$. We use two criteria, tracking success rate and location error for quantitative evaluations.

3.2 Experiment Results and Comparative Analysis

The visual evaluation of the comparative tracking results and quantitative location error are listed in Fig. 4 and Fig. 5, respectively. The success rates of all trackers are listed in Table 1.

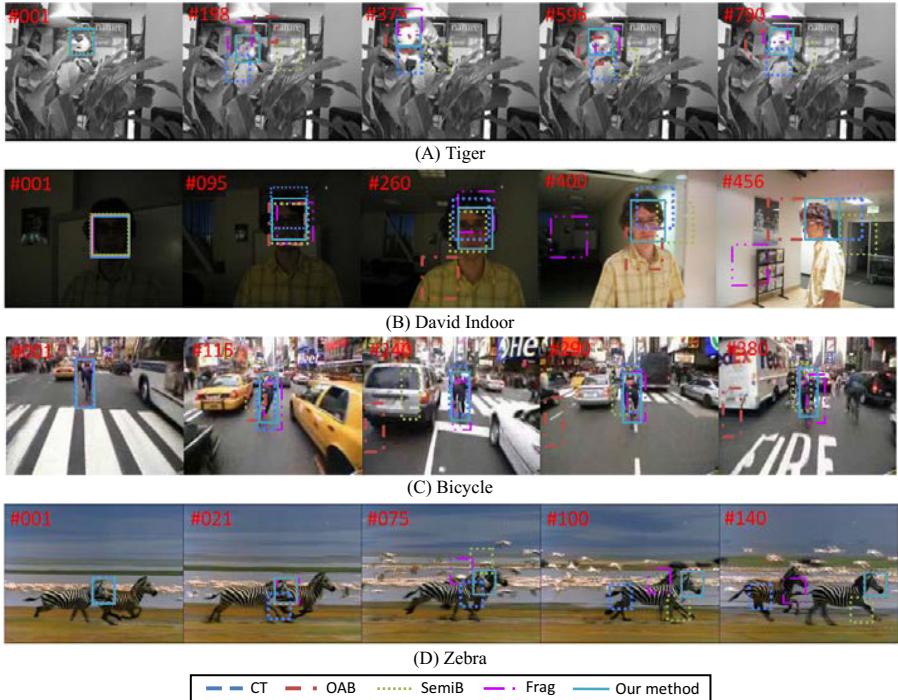


Fig. 4. Tracking result on four challenging sequences

Results show that our algorithm outperforms the others in tracking accuracy and robustness. CT performs well in *David Indoor* and *Bicycle* as it is able to deal objects with in-plane and out-of-plane rotation, but when there are severe occlusion and background clutter, it will lose objects. Table 1 confirms this conclusion, CT gets low success rates in sequence *Tiger* and *Zebra*. Frag loses the objects when the object undergoes both occlusion and rotation (e.g., *Tiger*), even if it is designed to handle partial occlusion. OAB will fail when object undergoes large pose change in a cluttered background (e.g., *Bicycle*). SemiB misses the object in many frames of sequences because of its semi-supervised procedure, therefore shows sharp fluctuations during tracking. The proposed algorithm runs at 50 frames per second on Pentium Dual-Core 3GHz with 4 GB RAM, which indicates that our tracking method has real-time performance and can cope with fast video object with relatively high frame rate.

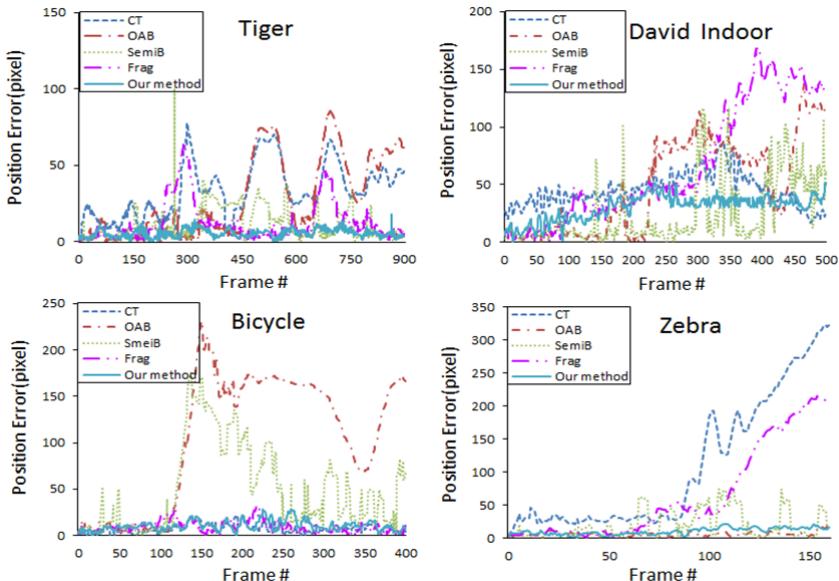


Fig. 5. The center location error of all test sequences

Table 1. Success rates (%). **Bold** mean the best value of success rate while *italic* indicates the second ones. The total number of evaluated frames is 3000.

	<i>Tiger</i>	<i>David</i>	<i>Bicycle</i>	<i>Zebra</i>
CT	5	34	64	13
OAB	29	37	7	73
SemiB	49	57	25	50
Frag	43	28	63	30
Our method	61	44	64	69

4 Conclusions

In this paper, an efficient tracking method is designed by using Gaussian kernel function to reweight feature, in which the object can be represented with much less dimension more sufficiently. Combined adaptive feature space regulation strategy with particle filter framework, the running time of algorithm was decreased without degrading the tracking accuracy. Some comparative experiments have been done to verify the effectiveness of our method.

There are many interesting ways to extend this work in future. The robustness of tracker can be improved with more sophisticated training mechanism of classifier, such as multiple instance learning as in [16]. Furthermore, it would be more meaningful to extend our system to be scale adaptive, which could extend the application areas.

Acknowledgments. This work was partly supported by the National Natural Science Foundation of China (No. 60875072, 61273350) and Beijing Natural Science Foundation (Grant 4112035).

References

1. Black, M.J., Jepson, A.D.: Eigentracking: Robust Matching and Tracking of Articulated Objects using a View-based Representation. *IJCV* 26, 63–84 (1998)
2. Lim, J., Ross, D., Lin, R., Yang, M.: Incremental Learning for Visual Tracking. In: Conference on Neural Information Processing Systems, pp. 793–800. MIT Press, Vancouver
3. Mei, X., Ling, H.: Robust Visual Tracking and Vehicle Classification via Sparse Representation. *PAMI*, 2259–2272 (2011)
4. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via online boosting. *BMVC*, 47–56 (2000)
5. Babenko, B., Ming-Hsuan, Y., Belongie, S.: Visual Tracking with Online MultipleInstance Learning. In: IEEE Conference on CVPR, pp. 983–990. IEEE Press, Maimi (2009)
6. Zhang, K., Zhang, L., Yang, M.-H.: Real-Time Compressive Tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part III. LNCS*, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
7. Bingham, E., Mannila, H.: Random Projection in Dimensionality Reduction: Applications to Image and Text Data. In: ACM International Conference on Knowledge Discovery and Data Mining, pp. 245–250. ACM Press, New York (2001)
8. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipshitz Mapping into Hilbert Space. In: Conference in Morden Analysis and Probability, vol. 26, pp. 189–206 (1984)
9. Ramesh, V., Meer, P.: Kernel-based Object Tracking. *PAMI*, 564–577 (2003)
10. Qing, W., Feng, C., Wenli, X., Ming-Hsuan, Y.: An Experimental Comparison of Online Object Tracking Algorithms. In: Proceedings of SPIE, Sandiego (2011)
11. University of Surrey, <http://libweb.surrey.ac.uk/library/skills>
12. Ross, D.A., Lim, J., Reuisung, L., Minghsuan, Y.: Incremental Learning for Robust Visual Tracking. *IJCV* 77, 125–141 (2008)
13. Ming Yang' Homepage,
<http://users.eecs.northwestern.edu/~mya671/VehicleVideo>
14. Helmut, Grabner' Homepage, <http://www.vision.ee.ethz.ch/~hegrabne/>
15. Adam, A., Rivlin, E., Shimshoni, I.: Robust Fragments-based Tracking using the Integral Histogram. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 798–805. IEEE Computer Society, Washington (2006)
16. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-Instance Learning via Embedded Instance Selection. *PAMI* 28, 1931–1947 (2006)

Harmonious Competition Learning for Gaussian Mixtures

GuoJun Liu and XiangLong Tang

Harbin Institute of Technology, China

Abstract. This paper proposes a novel automatic model selection algorithm for learning Gaussian mixtures. Unlike EM, we shall further increase the negative entropy of the posterior of latent variables to exert an indirect effect on model selection. The increase of negative entropy can be interpreted as a competition, which corresponds to an annihilation of those components with insufficient data to support. More importantly, this competition only depends on the data itself. Additionally, we seamlessly integrate parameter estimation and model selection into a single algorithm, which can be applied to any kind of parametric mixture model solved by an EM algorithm. Experiments involving Gaussian mixtures show the efficiency of our approach on model selection.

Keywords: Harmonious competition learning, Gaussian mixture model, Model selection, Expectation maximization.

1 Introduction

Gaussian mixtures as a flexible probabilistic modeling tool play an important role in many fields, such as machine learning, pattern recognition, bioinformatics, computer vision, signal and image analysis. Typically, Gaussian mixtures consists of K components. Supposed that each observation has been produced by exactly one of K components, to identify Gaussian mixtures, three levels of inference need to be solved, inferring which component produce each observation, i.e., inferring the parameters of each one of K components, and inferring the number of components, i.e., the value of K . The former two lead to a clustering of the set of observations, the last one is an important issue, also known as model selection or model comparison, which assigns a preference to a set of alternative statistical models with differing complexities. However, until now, there is little agreement on what on earth the best approach of model selection is.

Technically, the underlying mixture model is often not the one that fits the data best due to over-fitting. When the number of components K is fixed, maximum likelihood (ML) has proven to be an effective method of parameter estimation [1]. Nevertheless, if the value of K itself also needs to be estimated, maximum likelihood tends to be greedy and results in those over-parameterized models.

In this several decades, a great number of model selection methods have been proposed to avoid over-fitting, these methods can be broadly divided into four categories.

First, some methods attempt to indirectly compensate for the loss of the upper relation $\mathcal{M}_k \rightarrow \boldsymbol{\theta}$ by the addition of a penalty term $\mathcal{P}(\mathcal{M}_k)$ to the best-fit loglikelihood $\log p(\mathbf{X}|\boldsymbol{\theta}_{\text{ML}})$, such as cross-validation (CV) based criteria and the Akaike information criterion (AIC) [2]. Second, a constraint on the relation $\mathcal{M}_k \rightarrow \boldsymbol{\theta}$ is directly introduced by choosing a reasonable prior $p(\boldsymbol{\theta}|\mathcal{M}_k)$, the goal is to maximize $\log p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}_k)$, which is called maximum a posteriori (MAP) estimation in Bayesian approach. Similarly, in devising two-part coding schemes, both minimum description length (MDL) and minimum message length (MML) employ different approaches of parameter truncation to such a Bayesian situation. Third, the primary aim is to maximize the log marginal likelihood $\log p(\mathbf{X}|\mathcal{M}_k) = \log \int p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}_k) d\boldsymbol{\theta}$ by integrating out nuisance parameters. Unfortunately, in many cases, this Bayesian integral is generally difficult to compute, therefore, we have to resort to approximation schemes, such as Laplace's method used in Bayesian information criterion (BIC) [3], and variational approximation employed in variational Bayes (VB) [4]. Last, the competitive learning methods [5,6] have attracted more and more attentions for the ability of simultaneously dealing with both the parameter estimation and model selection. A important feature is that it can automatically perform component annihilation [7] , that is to say, the too weak component unsupported by data is simply annihilated by an explicit or heuristic competitive learning rule. However, there are still some problems and limitations for above methods.

In this paper, a novel automatic model selection algorithm is proposed to learn Gaussian mixtures. Unlike EM, we shall further increase the negative entropy of the posterior of latent variables to exert an indirect effect on model selection. The increase of negative entropy is virtually a transition from disorder to order, and also be interpreted as a competition. More importantly, this competition only depends on the data itself.

The rest of paper is organized as follows: in Section 2, we derive the harmonious competition learning on the basis of EM. In Section 3, we give a more detailed solution of harmonious competition function as a constrained optimization problem as well as its important properties. Section 4 reports experimental results on model selection for Gaussian mixtures and Section 5 ends the paper by presenting some concluding remarks.

2 Derivation of Harmonious Competition Learning Based on EM

In this section, we derive the harmonious competition learning on the basis of EM [8]. The EM algorithm [9] is an elegant and powerful technique to find maximum likelihood solutions for probabilistic models with latent nuisance variables \mathbf{Z} , it is an iterative optimization method to estimate some unknown parameters $\boldsymbol{\theta}$, in the light of the observed variables \mathbf{X} . The goal is to maximize the posterior probability of the parameters $\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{X})$. Equivalently, we can maximize the logarithm of the joint distribution which is proportional to the posterior:

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) \quad (1)$$

However, maximizing Eq. (1) inevitably involves the logarithm of a sum, which is difficult to deal with. Fortunately, by the Jensen's inequality, we can construct a tractable lower bound $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}}) \triangleq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q(\mathbf{Z})}$ in order to simply transform the log of a sum into a sum of logs $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}}) \leq \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q(\mathbf{Z})}$ where $q(\mathbf{Z})$ is an arbitrary probability distribution over the space of latent variables \mathbf{Z} .

In E-step, the optimal bound at a guess $\boldsymbol{\theta}^{\text{old}}$ can be obtained by maximizing $B(\boldsymbol{\theta}^{\text{old}}; \boldsymbol{\theta}^{\text{old}})$ with respect to the distribution $q(\mathbf{Z})$. Meanwhile, introducing a Lagrange multiplier λ to enforce the constraint $\sum_{\mathbf{Z}} q(\mathbf{Z}) = 1$, so we obtain

$$q(\mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}^{\text{old}})}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}^{\text{old}})} = p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \quad (2)$$

Subsequently, in M-step, we require to maximize $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\theta}$, and rewrite it as

$$\begin{aligned} B(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}}) &\triangleq \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})] + H_q \\ &= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] + \log p(\boldsymbol{\theta}) + H_q \end{aligned} \quad (3)$$

where $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to the distribution of $q(\mathbf{Z})$, $p(\boldsymbol{\theta})$ is the prior of the parameters $\boldsymbol{\theta}$, and H_q is the entropy of the distribution of $q(\mathbf{Z})$.

Generally, after the E-step, EM algorithm would fix $q(\mathbf{Z})$ at the value of $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ as Eq. (2), thereby the entropy H_q does not depend on $\boldsymbol{\theta}$. Maximizing the bound $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\theta}$ is up to the first two terms only:

$$\boldsymbol{\theta}^{\text{new}} = \arg \max_{\boldsymbol{\theta}} \{ \mathcal{Q}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \} \quad (4)$$

In particular, we must pay more attention to the first term $\mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})]$ in Eq. (3) rewritten as $\mathcal{Q}(q, \boldsymbol{\theta})$ by us, instead of $\mathcal{Q}(\boldsymbol{\theta})$ as the convention of EM. Note that $\mathcal{Q}(q, \boldsymbol{\theta})$ is not only a function of the parameters $\boldsymbol{\theta}$, but also a functional of the distribution $q(\mathbf{Z})$, which means that we can further tune the $q(\mathbf{Z})$ on the basis of the fixed value $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ in order to increase $\mathcal{Q}(q, \boldsymbol{\theta})$ before the change of the parameters $\boldsymbol{\theta}$ in M-step.

Therefore, a plug-in step, called harmonious competition step or C-step, is able to be inserted between E-step and M-step. In this step, the parameters $\boldsymbol{\theta}$ is still kept at the fixed value $\boldsymbol{\theta}^{\text{old}}$, then we have

$$\begin{aligned} \mathcal{Q}\left(q, \boldsymbol{\theta}^{\text{old}}\right) &= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}^{\text{old}})] + \log p(\boldsymbol{\theta}^{\text{old}}) \\ &= \mathbb{E}_q [\log p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}})] + \log p(\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \end{aligned} \quad (5)$$

Increasing $\mathcal{Q}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to $q(\mathbf{Z})$ only depends on the first term in Eq. (5), this is equivalent to further increase the negative entropy of $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$. It leads to

a new distribution $\hat{q}(\mathbf{Z}) = \mathcal{C}(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}))$ to ensure that $\mathcal{Q}(\hat{q}, \boldsymbol{\theta}^{\text{old}}) \geq \mathcal{Q}(q, \boldsymbol{\theta}^{\text{old}})$, that is to say,

$$\mathbb{E}_{\hat{q}} [\log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})] \geq \mathbb{E}_q [\log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})] \quad (6)$$

where $\mathcal{C}(\cdot)$ denotes harmonious competition function which is considered as a constrained optimization detailed in Section 3.3.

Last but not least, after C-step, \hat{q} as a new responsibility has taken the place of the old one $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ in M-step, then update and get new parameters $\boldsymbol{\theta}^{\text{new}}$. Note that the mixture weight as a subset of $\boldsymbol{\theta}^{\text{new}}$ has been updated by using \hat{q} instead of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, in other words, the result of harmonious competition has produced an effect on the mixture weight. More importantly, the increase of negative entropy is able to force the mixture weight of some components to tend to 0. Subsequently, we require another step called component annihilation. In this step, annihilate one or more components whose mixture weight is less than the predefined threshold ϵ , then remove the corresponding parameters from the set of parameters and normalize the mixture weights once more, at the same time, update K to be the number of the survived components. i.e., automatic model selection.

3 Harmonious Competition Learning

3.1 The Relation of Negative Entropy and Competition

Negative entropy is viewed as a mathematical synonym for *order* in an entropic sense, this term comes from Nobel laureate Erwin Schrödinger's famous booklet *What is life?*.

For a probability distribution, with the increase of negative entropy, it will transition gradually from disorder or chaos to order. Geometrically, it can be interpreted as a collapse from a high-dimensional space to a lower dimensional subspace. For example, suppose that a change of a probability distribution with 3 elements like that $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \rightarrow (\frac{1}{2}, \frac{1}{2}, 0) \rightarrow (1, 0, 0)$, the corresponding change of negative entropy is $-\log 3 < -\log 2 < 0$. Once the value of some element is equal to 0, in geometry, it means that the corresponding spatial dimension plays no role in describing the current probability distribution. Consequently, it forms a collapse into a subspace.

3.2 The Probability Simplex

Simplex is an important family of polyhedra. Specifically, a $(n-1)$ -dimensional simplex is the convex hull of its n vertices, e.g., a 0-dimensional simplex is a single point, a 1-dimensional simplex is a line segment, a 2-dimensional simplex is a triangle, and a 3-dimensional simplex is a tetrahedron.

3.3 Harmonious Competition Function

The negative entropy of a discrete probability distribution is defined by $h(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i$, where $\mathbf{x} \in \mathbb{S}^n$. To increase negative entropy $h(\mathbf{x})$, the gradient based method is available to get \mathbf{g} as following

$$g_i = x_i + \eta \nabla_{x_i} h = x_i + \eta (1 + \log x_i) \quad (7)$$

where $\eta > 0$, η is a learning rate and also called a competition intensity. Note that the vector \mathbf{g} may be not in the probability simplex any more, i.e., $\mathbf{g} \notin \mathbb{S}^n$. Therefore, to turn it into a probability distribution, we just need to project it onto the probability simplex and find its corresponding projection $\mathbf{y} \in \mathbb{S}^n$. This is equivalent to solve a convex optimization problem, we consider it in the standard form.

$$\begin{aligned} \text{minimize} \quad f_0(\mathbf{y}) &= \frac{1}{2} \sum_{i=1}^n (y_i - g_i)^2 \\ \text{subject to} \quad \mathbf{y} &\succeq \mathbf{0}, \quad \mathbf{1}^\top \mathbf{y} = 1 \end{aligned} \quad (8)$$

Introducing Lagrange multipliers $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ for the inequality constraints $\mathbf{y}^* \succeq \mathbf{0}$ and a multiplier $\nu^* \in \mathbb{R}$ for the equality constraint $\mathbf{1}^\top \mathbf{y} = 1$, We define the Lagrangian \mathcal{L} associated with the problem as

$$\mathcal{L}(\mathbf{y}, \boldsymbol{\lambda}^*, \nu^*) = \frac{1}{2} \sum_{i=1}^n (y_i - g_i)^2 + \sum_{i=1}^n \lambda_i^* (-y_i^*) - \sum_{i=1}^n \nu^* y_i^* \quad (9)$$

These above equations satisfy the KKT conditions and can be solved directly to find \mathbf{y}^* , $\boldsymbol{\lambda}^*$, and ν^* . Thus we have

$$y_i^* = \begin{cases} \nu^* + g_i & \nu^* > -g_i \\ 0 & \nu^* \leq -g_i \end{cases} \quad (10)$$

or, put more simply, $y_i^* = \max \{0, \nu^* + g_i\}$. Substituting it into the second condition $\mathbf{1}^\top \mathbf{y}^* = 1$, we obtain

$$\sum_{i=1}^n \max \{0, \nu^* + g_i\} = 1 \quad (11)$$

This solution method is called water-filling. The left-hand side is a piecewise-linear increasing function of ν^* , with breakpoints at $-g_i$. Therefore, it is solvable and has a unique solution.

For convenience, we shall give a definition of the above method and call it harmonious competition function.

Definition 1 (Harmonious competition function). Let \mathbf{x} and \mathbf{y} be two n -dimensional vectors in the probability simplex, i.e., $\mathbf{x}, \mathbf{y} \in \mathbb{S}^n$, the harmonious competition function $\mathcal{C} : \mathbf{x} \mapsto \mathbf{y}$ is defined by

$$y_i = \mathcal{C}(x_i) = \max \{0, x_i + \Delta_i + v\}, \quad i \in \{1, \dots, n\} \quad (12)$$

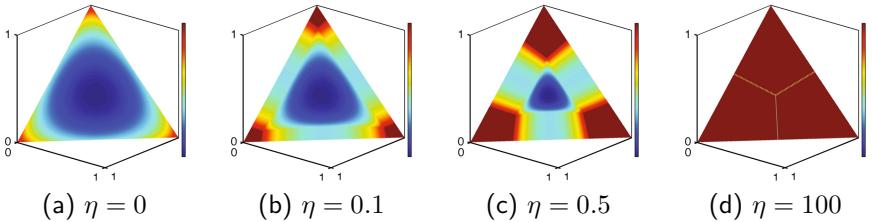


Fig. 1. The change of negative entropy before and after the harmonious competition function. The value of negative entropy ranges from $-\log 3$ to 0, correspondingly, it is described by the jet colormap which ranges from blue to red, and passes through the colors cyan, yellow, and orange.

where $\Delta \stackrel{\text{def}}{=} f(\mathbf{x})$, f is a monotonically increasing function and $\Delta \in \mathbb{R}^n$. v is a single variable and chosen such that $\sum_{i=1}^n \max\{0, x_i + \Delta_i + v\} = 1$.

Subsequently, we shall give a quantitative analysis of how to increase negative entropy with the different value of competition intensity, as illustrated in Fig. 1. Suppose that $\mathbf{x} \in \mathbb{S}^3$, then calculate $h(\mathbf{x})$, as shown in Fig. 1(a), where $\eta = 0$ such that $h(\mathbf{y}) = h(\mathbf{x})$ by Eq. (7). Let $\mathbf{y} = \mathcal{C}(\mathbf{x})$ for every \mathbf{x} , and redraw $h(\mathbf{y})$ onto the same probability simplex with different η , as illustrated in Fig. 1(b)-(d). Here, η governs the intensity of the competition. For example, in Fig. 1(d), for almost all vectors in the domain, the value of negative entropy is approximately equal to 0, it means that a competition is so intense to make the probability of one element equal to 1 and the probability of the other two elements equal to 0. In other words, the competition reassigns the probability of each element. When η tends to infinity, harmonious competition will degenerate to a winner-take-all manner of K -means.

4 Experiments

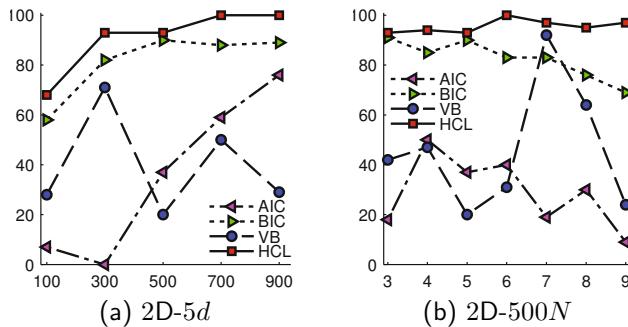
Although the proposed method can be applied for any kind of mixture model, our experiments focus only on Gaussian mixtures, which are by far the most common model. To compare our algorithm (HCL) with those traditional methods referred in Section 1, we chose AIC, BIC and VB, as the most commonly used model selection criterion. For all experiments, we set the candidate model of AIC and BIC range from $K_{min} = 1$ to $K_{max} = 10$, where K denotes the number of mixture components. In addition, the Dirichlet distribution as the conjugate prior of multinomial distribution is often used in VB, the value of its hyperparameter acts as a prior knowledge and has an important effect on model selection, therefore, let it equal to a small value, i.e., an uninformative prior. Last, we set the initial number of mixture components K_{init} be large enough, e.g., $K_{init} = 15$ in VB and our method.

There are two groups of experiments, for every data set, all methods run 100 times respectively, then compare the results with the true number of mixtures components, and obtain the percentage of success of various methods.

Table 1. Percentage of success of various methods of two real data sets

DATA SET	AIC	BIC	VB	HCL
Old faithful	0	100	93	100
Iris data	0	2	33	65

In the first example, we consider the two well-known real data sets, one is Old Faithful, a 272 2-dimensional bimodal data set, the other is Iris data set, 150 4-dimensional points from three classes, 50 per class. The results are shown in Table 1. For a large sample low-dimensional data set, BIC, VB and our method have a good performance. Notwithstanding a sharp decline in the correct model selections with the dimensional increase and the decrease of sample number, our approach is still superior to other methods.

**Fig. 2.** Percentage of success of various methods using 2-dimensional synthetic data with different N and d respectively

Second, we use N samples from a 5-component bivariate mixture, the mixture weight of each component is equal to $1/5$, mean vectors at $[0, 0]^T$, $[0, d]^T$, $[0, -d]^T$, $[d, 0]^T$, $[-d, 0]^T$ where d denotes the distance from the origin, and equal covariance matrices $\text{diag}\{2, 0.2\}$. In Fig. 2(a), we fix $d = 5$ and draw different number of samples from above Gaussian mixtures, N ranges from 100 to 900. Next, we fix $N = 500$, then use different d to generate samples.

As illustrated in Fig. 2, the performance of those traditional methods is unstable, worse for most cases and better only for some special cases which seem suitable for necessary approximation conditions, such as AIC, BIC. As for VB, it naturally embodies many features of Bayesian inference, so it automatically makes the trade-off between fitting the data and model complexity, but most importantly, that which one is more comprise-inclined in practice is not clear. In contrast, the harmonious competition only depends on data itself, instead of some approximation techniques or heuristic rules, therefore, our method is more robust.

5 Conclusion

This paper proposes a novel automatic model selection algorithm for learning Gaussian mixtures. The novelty in our approach is that harmonious competition is able to make the mixture weight of those components with insufficient data to support tend to zero, more importantly, it only depends on the data itself. Furthermore, we seamlessly integrate parameter estimation and model selection into a single algorithm, which can be applied to any kind of parametric mixture model solved by an EM algorithm. Experiments involving Gaussian mixtures show the efficiency of our approach on model selection.

Acknowledgments. This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. HIT.NSRIF.2014069) and National Natural Science Foundation of China (Grant No. 61173087).

References

1. Lanterman, A.D.: Schwarz, wallace, and rissanen: Intertwining themes in theories of model selection. *International Statistical Review* 69(2), 185–212 (2001)
2. Akaike, H.: A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723 (1974)
3. Schwarz, G.: Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464 (1978)
4. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. *Machine Learning* 37(2), 183–233 (1999)
5. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. *IEEE Transactions on Neural Networks* 4(4), 636–649 (1993)
6. Xu, L.: Bayesian Ying-Yang system, best harmony learning, and five action circling. *Frontiers of Electrical and Electronic Engineering in China* 5(3), 281–328 (2010)
7. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(3), 381–396 (2002)
8. Neal, R.M., Hinton, G.E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M.I. (ed.) *Learning in Graphical Models*, 1st edn., pp. 355–368. MIT Press, Cambridge (1998)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1), 1–38 (1977)

A Hierarchical Path Planning Approach Based on Reinforcement Learning for Mobile Robots

Qi Guo, Lei Zuo, Rui Zheng, and Xin Xu

College of Mechatronics and Automation, National University of Defense Technology,
Changsha 410073, China
guoqics@gmail.com

Abstract. In this paper, we propose a novel hierarchical path planning algorithm for mobile robots based on A* and reinforcement learning (RL) with the structure of two layers. In the first layer, we adopt the A* search algorithm to plan a geometric path and select several points as sub-target points for the planning of the next stage. In the second layer, a local path planning algorithm based on an approximate RL method called Least Square Policy Iteration (LSPI) is used to find a kinematically feasible path with these sub-targets. After learning, the local path planner in the second layer has good generalization performance. The path obtained by the proposed algorithm is smooth and safe for executing. Simulations have been carried out and the results demonstrate the validity of the proposed scheme.

Keywords: Autonomous mobile robots, path planning, Least Square Policy Iteration, A* search, Hierarchical planning.

1 Introduction

Path planning generally includes local path planning and global path planning. Until now, various local path planning algorithms have been proposed, e.g. path planning based on simulated annealing, the potential field approach, the rolling window algorithm, path planning using artificial neural networks, ant colony optimization, particle swarm optimization, the A* algorithm, and the D* algorithm, etc. Common global path planning algorithms include: grid map, topology graph, visibility graph, Voronoi graph, genetic algorithms and the level-set algorithm [1], etc.

The A* algorithm is a heuristic search algorithm that has been popularly studied. However, with abrupt changes of direction and clinging to obstacles, the path may not be feasible to be executed by real mobile robots. The Potential Field method was proposed in 1986 by Khatib and Krogh. Since the Potential Field method is based on local information, the goal can't be reached in some cases. Probabilistic sampling-based methods for path planning, like the probabilistic roadmap (PRM) [6], rapidly exploring random trees (RRT) [5], and expansive space trees (EST) [7], were shown to be particularly useful when the configuration space of the moving body has a large number of degrees of freedom (DOFs). Nevertheless, without taking robot's

kinematics or dynamics constraints into consideration, path planning algorithms will usually generate non-smooth or even kinematically infeasible paths. So, much effort and modifications are needed to apply those algorithms to real robots [2].

With the development of computational intelligence (CI), there are more and more research interests on path planning methods based on CI [3, 4]. For example, fuzzy logic has been used for the path planning problem in dynamic and uncertain environments. Er, et al. [9] presented a hybrid learning approach for obstacle avoidance of a mobile robot, where a neuro-fuzzy controller was developed from a pre-wired or innate controller based on supervised learning in a simulation environment. As a class of global search methods, the Genetic Algorithms (GAs) have also been widely studied to generate near-optimal paths by taking advantages of its strong optimization ability. However, conventional random mutation operator in simple GAs or some other improved mutation operators can cause infeasible paths. A new mutation operator was proposed in [10] for GAs and applied to the path planning problem of mobile robots in dynamic environments.

Reinforcement learning (RL) is a machine learning framework for sequential decision making under uncertainties [8]. The environment of RL is typically modeled as a Markov decision process. As we will discuss later, the path planning problem for autonomous mobile robots can be modeled as a Markov Decision Process (MDP). Due to the optimization and generalization ability of reinforcement learning, it is very promising to apply RL in path planning problems.

As a popularly studied RL algorithm, Q-learning deals with MDPs with discrete state and action spaces. However, when the state and action spaces become large or continuous, the computational complexity of Q-learning increases exponentially. That is so called the “curse of dimensionality”. To solve this problem, the research on RL has been focused on value function approximation and policy approximation for MDPs with large or continuous spaces. In [11], Lagoudskis and Parr proposed the Least-squares Policy Iteration (LSPI) algorithm which has been shown to have good convergence and generalization ability.

In order to solve the above mentioned problem in traditional path planning algorithms for mobile robots, this paper proposed a new hierarchical path planning approach which includes a higher level planner and a lower level planner. For the higher level planner, we adopt the A* algorithm to search a short, obstacle-free geometric path. For the lower level, we sample the state space with the kinematics constraints of mobile robots, and use LSPI to learn a smooth path planner. The learned planner is a reactive navigation policy considering the kinematics constraints of mobile robots and it has good generalization performance. Due to the generalization ability of LSPI, the lower level planner is adaptive to new environments without re-learning. What's more, the planning results are kinematically feasible.

The rest of this paper is organized as follows: in section 2, we introduce the whole framework of the proposed algorithm, and the lower level planner based on LSPI. In section 3, we discuss the implementations and performance evaluations of the proposed path planning method, as well as comparisons with other path planners. Section 4 concludes with a summary of our contributions.

2 The A*-LSPI Algorithm for Hierarchical Path Planning

2.1 Description of the Mobile Robot and the MDP Model

The state of the robot is defined as $[x_t, y_t, \theta_t]$ under the global coordinate, where x_t and y_t are the robot's position and θ_t is the angle between the forward direction of the robot and horizontal axis. The mobile robot has an omni-directional wheel and two driving wheels. The angular speeds of the driving wheels can be controlled. Six ultrasonic sensors are equipped in the front part of the robot. The detection distance of each sensor is d , and the detection angles are 30 degrees.

The robot's path which is a sequence of the robot's states has the Markovian property. We can model the process of path planning as two MDPs. One MDP is used to describe the behavior of approaching the goal and the other is to model the process of obstacle avoidance. During simulations, the state, action, and reward of the MDP can be defined in Table 1:

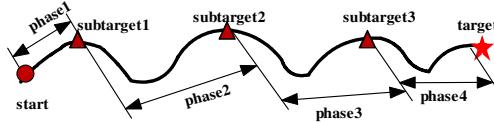
Table 1. Definition of the MDP models

	The MDP model of approaching the goal	The MDP model of learning to avoid obstacles
state	$[d_g, \phi_g]$	(S1,S2, S3,S4, S5,S6)
action	The combination of the speed of left and right wheel: forward: [0.5, 0.5] turn right: [0, 0.5] turn left: [0.5, 0]	The combination of the speeds of the left and right wheel: forward: [0.5, 0.5] turn right: [0, 0.5] turn left: [0.5, 0]
reward	$\begin{cases} 1, & \text{if the robot achieve the goal} \\ -d_g - \phi_g, & \text{else} \end{cases}$	See Table 2, where $k=0.9$

2.2 The Framework of the Proposed Algorithm

We use the A* algorithm as the higher level planner, and sampling the result of a geometric path as the *subtarget* points of the lower level planner. Fig. 2 outlines the process of the path planning method. During each phase, we use the lower level planner to generate a sub-path between two neighboring sub-targets. The combination of sub-path is the ultimate planning result.

The advantage of this hierarchical idea is that either planner may be developed independently of the other: The discrete planner maybe tailored to satisfy vehicular task specifications (finding low-cost, obstacle-free paths is an example of such a task), whereas trajectory planning schemes that are based on control theory may be tailored to cope with complex vehicle dynamics.

**Fig. 1.** The process of lower lever path planning

2.3 Path Planner Based on LSPI in the Lower Level

For path planner based on LSPI, we separate the process as two sub-problems: learning to approach the target and learning to avoid obstacles.

(1) Learning to Approach the Target Point

During the process of approaching the target, the robot's next states are determined by the current position and the action it takes. The state vector of the MDP includes the distance between the mobile robot and target d_g , and the angel between the direction of mobile and the target $\phi_g \in [0, \pi]$. When the mobile robot arrives at the target point, the immediate reward is set as 1, otherwise, the reward is set as $-\alpha \cdot d_g - \beta \cdot \phi_g$, where α and β are two constants.

(2) Learning to Avoid Obstacles

During the process of avoiding obstacles, the robot's next states are also determined by the current position and the action it adopts. So, the process can be modeled as another MDP. The detailed reward functions are set as follows:

Table 2. The reward function of avoiding obstacles

<i>Series numble</i>	<i>conditions</i>	<i>illuminate</i>	<i>Reward function</i>
1	$d_{\min} \leq D_{us}$	The smallest reading of sensors less than or equal to a certain safety valve	-10
2	$d_l \leq D$ $d_r \leq D$ $d_{\min} > D_{us}$	The average readings of left and right sensors are both less than safe distance of avoiding obstacles	$-k(D-d_l)-k(D-d_r)$
3	$d_l \leq D$ $d_r > D$ $d_{\min} > D_{us}$	The average reading of left sensors is less than safe distance of avoiding obstacles	$-k(D-d_l)$
4	$d_l > D$ $d_r \leq D$ $d_{\min} > D_{us}$	The average reading of right sensors is less than safe distance of avoiding obstacles	$-k(D-d_r)$
5	$d_l > D$ $d_r > D$ $d_{\min} > D_{us}$	The average reading of left and right sensors are both larger than safe distance of avoiding obstacles	10

Let S_i denote the reading of distance sensor i ($i=1,2,\dots,6$). In Table 2, $d_{\min} = \min(S_1, S_2, S_3, S_4, S_5, S_6)$, $d_l = \text{average}(S_1, S_2, S_3)$, $d_r = \text{average}(S_4, S_5, S_6)$, $k > 0$ is a proportional constant, D is the safe distance of avoiding obstacles, D_{us} is the distance for an emergency brake. The function is triggered when the reading of any sensor is less than the safety threshold, and the reward is -10.

During the process of sampling, the robot takes the readings of sensors as the current state. Then take a random action, after fixed number of steps, the robot takes the readings of every sensor as the next state, and receives the reward based on the above table. If the reading value of any sensor is less than the safety threshold, this sampling period will be over, and the next period will begin. After a sampling phase, the training phase based on LSPI will learn two policies of approaching the target point and avoiding obstacles, respectively. The whole flowchart of the proposed algorithm is as follows.

Table 3. The whole framework of the proposed algorithm

Algorithm 1: The hierarchical path planning algorithm

Level 1: A* search

Input: grid-based map Gm;
Start point S, and target point T;

Output: subtarget points ST;

Level 2: path planner based on LSPI

Sampling:

Output: Samples;

Training:

Input: Samples

Output: Approaching target policy FP

Avoiding obstacle policy AP

Planning:

For i = 1 : size (ST)

While position (robot) != ST (i)

If distance(robot, obstacles) < D (the boundary of switch planning policy)

 Use AP to generate an action

else

 Adopt FP to generate an action

End

 Update the robot's position

end

end

success, output the final path PATH

2.4 Least Square Policy Iteration (LSPI) for Lower-level Planning

The LSPI algorithm approximates the value functions of an MDP by observing data generated from the state transitions and the rewards of the MDP. The Bellman equation of an MDP can be expressed as matrix format: $Q^\pi = R + \gamma P \Pi_\pi Q^\pi$, where

$$R(s, a) = \sum_{s' \in S} P(s, a, s') R(s, a, s') \quad (1)$$

In LSPI, the state-action value function $Q^\pi(x, a)$ can be approximated using a linearly weighted combination of n basis functions:

$$\hat{Q}^\pi(x, a) = \vec{\phi}^T(x, a) W \quad (2)$$

where $W = (w_1, w_2, \dots, w_n)^T$ is the weight vector and $\vec{\phi}(x, a)$ is the basis function vector, denoted by:

$$\vec{\phi}(x, a) = (\phi_1(x, a), \phi_2(x, a), \dots, \phi_n(x, a))^T \quad (3)$$

Let

$$\Phi = \begin{pmatrix} \vec{\phi}^T(x_1, a_1) \\ \vec{\phi}^T(x_2, a_2) \\ \vdots \\ \vec{\phi}^T(x_m, a_m) \end{pmatrix}, \quad \Phi' = \begin{pmatrix} \vec{\phi}^T(x_1, a_1) \\ \vec{\phi}^T(x'_2, a_2) \\ \vdots \\ \vec{\phi}^T(x'_m, a_m) \end{pmatrix}, \quad R = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix}$$

where $D = \{(x_i, a_i, r_i, x'_i, a'_i) \mid i = 1, 2, \dots, m\}$ is a set of collected samples from an initial policy $\pi[0]$ and $a'_i \sim \pi[0](alx'_i)$.

Then, the least-squares fixed-point solution for action value function approximation [27] and the corresponding improved policy can be obtained as follows [13]:

$$\begin{cases} W^{\pi[t]} = (\Phi^T(\Phi - \gamma\Phi'))^{-1}\Phi^T R & t=0,1,\dots \\ \pi[t+1](x) = \arg \max_a \vec{\phi}^T(x, a) \omega^{\pi[t]} \end{cases} \quad (4)$$

By using the MDP models defined in Table 1, if enough samples are collected via simulation, the LSPI algorithm can be used to learn a path planning strategy which is composed of two policies. One policy is for approaching the goal and the other is used for avoiding obstacles. In the following, we will compare the performance between the pure LSPI-based lower level planner with the proposed hierarchical approach.

3 Simulations and Evaluations

In the simulation, we considered different environments with different sizes and different obstacles. The size and shape of all the obstacles were selected randomly, and the locations were also set randomly. The green triangle is the starting point, and the red triangle is the goal. After a lot of simulations, three typical sets of results are shown below to evaluate the proposed algorithm.

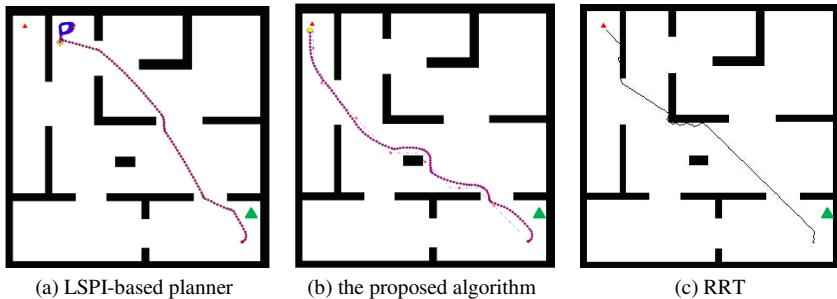
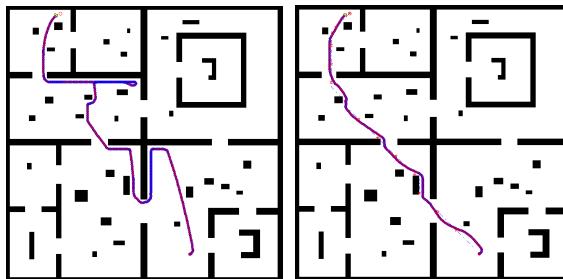


Fig. 2. The planning results of 3 different algorithms(The green triangle are the start point, the red triangle is the goal)

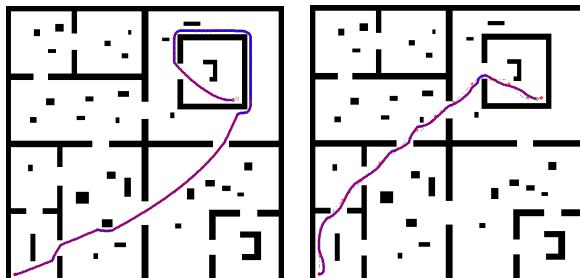
In the simulation, we compared the performance of three different path planners. The first one is a single lower level path planner based on LSPI. The second one is the proposed hierarchical path planner and the third is the Rapidly-exploring Random Tree (RRT) approach. The size of the map we adopted is 300×300 . From Fig. 2, we

find out that the LSPI-based lower level planner is difficult to find out the available path between the starting point and the goal point in some cases. In the proposed approach, we utilize the A* search on the first layer, and pass the sub-target points to the second layer. The red pentacles on the Fig. 2(b) are the sub-target points, and the blue dotted line is the result of A* search. The result of A* search is difficult for the mobile robot to execute. Fig. 2(c) is the result of RRT. It is also shown that the planning result of the proposed approach is smoother than that of RRT.

The following figure also illustrates that the advantages of the hierarchical planning framework. It is shown that the proposed hierarchical approach can obtain much better planning results than the pure LSPI-based lower level planner.



(a) The first set of experiment result. The left one is the result of the LSPI-based planner and the right one is the result of the proposed algorithm



(b) The second set of experiment result. The left one is the result of the LSPI-based planner and the right one is the result of the proposed algorithm

Fig. 3. Planning results of pure LSPI-based planner and the hierarchical approach

4 Conclusions

In this paper, a hierarchical path planning method called A*-LSPI was proposed. In the proposed method, we combine a global planning algorithm A* and a local RL-based planning algorithm based on LSPI together to get better planning results. In the first layer, we adopt the A* search algorithm for global planning and select several points on the obtained path as sub-target points of the second layer. In the second layer, we adopt the local path planning policy trained by the LSPI algorithm to re-plan

the path to these sub-target points. With the constraint of mobile robot's kinematics in the second layer, the path obtained by the proposed algorithm is kinematically feasible. What's more, the policy of the second layer has good generalization performance which is capable to re-adapt to new environments without re-sampling or re-learning. The simulation results have illustrated the effectiveness of the proposed method.

Acknowledgements. This paper is supported by National Natural Science Foundation of China under Grant 61075072, & 91220301, the Program for New Century Excellent Talents in University under Grant NCET-10-0901.

References

1. Lolla, T., Ueckermann, M.P., Yigit, K., Haley, P.J., Lermusiaux, P.: Path planning in time dependent flow fields using level set methods. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 166–173 (2012)
2. Cowlagi, R.V., Tsiotras, P.: Hierarchical Motion Planning With Dynamical Feasibility Guarantees for Mobile Robotic Vehicles. *IEEE Transactions on Robotics* 28(2), 379–396 (2012)
3. Cai, Z., Peng, Z.: Cooperative coevolutionary adaptive genetic algorithm in path planning of cooperative multi-mobile robot system. *Intelligent & Robotic System* 33(1), 61–67 (2002)
4. Nilsson, N.: A mobile automation: An application of artificial intelligence techniques. In: Proc. 1st Int. Joint Conf. Artificial Intelligence, Washington, DC, pp. 509–520 (1969)
5. Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Autom.* 12(4), 566–580 (1996)
6. LaValle, S.M., Kuffner, J.J.: Rapidly-exploring random trees: Progress and prospects. *Algorithmic and Computational Robotics: New Directions*, 293–308 (2001)
7. Hsu, D., Latombe, J., Motwani, R.: Path planning in expansive configuration spaces. *Int. J. Comp. Geo. Appl.* 4, 495–512 (1999)
8. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press (1998)
9. Er, M.J., Deng, C.: Obstacle Avoidance of a Mobile Robot Using Hybrid Learning Approach. *IEEE Transactions on Industrial Electronics* 52(3), 898–906 (2005)
10. Tuncer, A., Yildirim, M.: Dynamic path planning of mobile robots with improved genetic algorithm. *Computers and Electrical Engineering* (2012)
11. Lagoudakis, M.G., Parr, R.: Least-Squares Policy Iteration. *Journal of Machine Learning Research* 4, 1107–1149 (2003)

Color Image Segmentation Based-on SVM Using Mixed Features and Combined Kernel

Lei Li¹, Dong yan Shi², and Jun Xu³

¹ The Science in NUPT

lil@njupt.edu.cn

² Pattern Recognition and Intelligent System

dongdongssdy@163.com

³ Control Theory and Control Engineering

xjback@126.com

Abstract. Image segmentation occupies the important position in image processing, so both high-efficiency and accurate segmentation are of great importance to image's subsequent research. In this paper, taking the color and texture features into account, the color features are combined with RGB and HSV color space features, and the texture features are made up with the mean value and standard deviation. Considering the advantages of support vector machine (SVM) in classification, we convert the segmentation problem into a classification problem, and we can get the satisfied segmentation result. Genetic Algorithms (GA) is used for the optimization of the parameters in SVM's kernel which is combined with polynomial (poly) and radial basis function (rbf) kernel reasonably by mercer theorem. The experimental results show that we can segment various kinds of color images effectively.

Keywords: color image segmentation, svm, combined features, GA, combined kernel.

1 Introduction

Image segmentation is an important image analysis technique. In images' research and application, people are usually interested in some unique quality area of the image, which is always known as goal or foreground (the other part is called background). The unique quality can be the grey value of pixels, the profilogram of object, color, and texture and so on. By certain segmentation technology, we can get the part of images that we need, and that is the foreground part.

We can divide the foreground and the background of the image into two categories, for which we choose the part we need, so the segmentation can be converted into a classification problem. Support vector machines (SVM) have been successfully applied in classification and function estimation problems after their introduction by Vapnik within the context of statistical learning theory and structural risk minimization. Vapnik constructed the standard SVM to classify training data into two

classes [1] and because of SVM's great advantages in classification, it can be utilized successfully in image segmentation. SVM is a function estimation problem, based on the context of statistical learning theory and structural risk minimization [2]. The kernel function maps the low dimension space to high dimension space, and transforms non-linearity into linearity, so we can improve the image segmentation effectively by selecting suitable kernel function.

This paper considered the combination of the color and the texture of the image as training features, and the color space feature is the combination of RGB and HSV color space, while the texture feature is the combination of the mean value and standard deviation of entropy, inertia moment and the correlation. Take these 4 features (2 color space features and 2 texture features) as training features for SVM to get reasonable classification. The kernel function used in this paper is the combination kernel function and the segmentation result is satisfied in our experiment.

The article is arranged as follows: This section is about introduction, the second section is about the extraction of features, and then the description of SVM and kernel function, the last section is about the experimentation and the analysis of the experiment results.

2 Feature Extraction

Images used in this paper are grabbed from image library at the University of California Berkeley, and the most important features of the color image are color feature and texture feature.

2.1 Color Space Features

The most important color in the spectrum are three primary colors: red(R),green (G), blue(B). Because the image acquisition and display devices use the RGB color space, RGB color space is the most basic and the most commonly used color space in color image processing. Reference [3] shows that the RGB plays an important role in image segmentation. The feature vector can be expressed by : $T_{10} = \{R, G, B\}$.

Smith proposed the HSV color space, a color model indicating visual perception, including three continuous components: one is tone, it referring to the type of color; the second is saturation level, expressing the purity of color; and the third is light intensity showing the degree of light and shade. This feature space is more intuitive, as reference [4] makes good use of the color space for image processing. The feature vector is $T_{20} = \{H_1, S, V\}$.So, the total Color feature vector is $T_1 = \{T_{10}, T_{20}\}$.

2.2 Texture Features

Haralick proposed the famous gray level co-occurrence matrix (GLCM).AS a good method of analyzing texture feature, it is widely used in transforming gray value into texture information.Although the texture features extracted by gray level co-occurrence matrix (GLCM) have better resolution capability, it takes too much

time in calculating GLCM and extracting 14 texture features. reference[2] had a detailed study of 6 texture features and concluded that the most important feature among them were contrast ratio and entropy ;reference[5] analyzed the calculation problem of GLCM and obtained 3 best resolution and uncorrelated features: contrast ratio(inertia moment), entropy and the correlation. To save time, this paper set the distance as 1 and angle of 0, 45, 90, 135.

Using GLCM to express the joint probability distribution of two gray level of pixel which are at a distance of $(\Delta x, \Delta y)$. If the gray level of the image is L, then the co-occurrence matrix is $L \times L$, and it can be expressed as $M_{(\Delta x, \Delta y)}(h, k)$. The element m_{hk} located at (h, k) shows the occurrences of the pixel in gradation h and k $(\Delta x, \Delta y)$ apart. Texture features can be expressed through the use of the three following features.

(1) inertia moment(I) :

$$I = \sum_h \sum_k (h - k)^2 m_{hk} \quad (1)$$

For rough texture, the value of m_{hk} concentrated near the main diagonal relatively, and at this moment the value of $(h-k)$ is smaller, so the corresponding value of I is smaller too. In contrast, the corresponding value I is bigger in the fine texture.

(2) Entropy (H_2) :

$$H_2 = - \sum_h \sum_k m_{hk} \log m_{hk} \quad (2)$$

When m_{hk} in GLCM is invariant and is scattered, H will be larger; otherwise, the value of H will be much smaller when m_{hk} is concentrated.

(3) the correlation(C) :

$$C = [\sum_h \sum_k h k m_{hk} - \bar{x}_x \bar{x}_y] / \sigma_x \sigma_y . \quad (3)$$

$\bar{x}_x, \bar{x}_y, \sigma_x, \sigma_y$ are the mean value and standard deviation of m_x, m_y ; $m_x = \sum_k m_{hk}$ is the sum of each column elements of matrix M; $m_y = \sum_h m_{hk}$ is the sum of each row elements of matrix M. The correlation is used to describe the similarity between row and column elements of the matrix, and it is a measure of gray linear relationship.

Calculate the mean value and the standard deviation of entropy, inertia moment and correlation, and the result is expressed by $I_m, H_{2m}, C_m, I_s, H_{2s}, C_s$.

The image's texture feature vector $T_2 = \{I_m, H_{2m}, C_m, I_s, H_{2s}, C_s\}$ is made up by above 6 features.

2.3 Feature Normalization

Because the features are not in an order of magnitude, the normalization is necessary. In this paper, we used linear function transformation, and the expression is as follows: $y = (x - \text{MinValue}) / (\text{MaxValue} - \text{MinValue})$, x and y respectively represent the values before and after the transformation, maximum and the minimum are the biggest and the smallest value in each column sample features. After the normalization, the feature vector will be bound in [-1, 1]. Merge each column after normalization to assemble into a feature ,and then combine all the features.

So, the final feature vector is

$$T = \{R_n, G_n, B_n, H_{1n}, S_n, V_n, I_{mn}, H_{2mn}, C_{mn}, I_{sn}, H_{2sn}, C_{sn}\}$$

R_n stands for the normalized feature R.

The normalization is not only used in the train features, but also used in the test features of the image we need to segment.

3 SVM and Kernel Function

3.1 The Theory of SVM

The main ideas of SVM can be summarized as two points: the linearly separable situation and the non-linear problem.Such as literature[6]: when linearly separable, we select the training set and the appropriate parameters, construct and solve the convex quadratic programming , and finally separate the samples by classification decision function; when linearly non-separable, the training set will be transformed from original feature space to Hilbert space, that is, transforming the non-linear problem in low-dimensional space into linearly separable problem in high-dimensional space, the subsequent steps are the same as linearly separable situation.

3.2 Kernel Function and Hybrid Kernel Function

The advantage of SVM is mainly reflected in solving liner non-separable problem by introducing kernel function. It ingeniously solved the inner product operation in high-dimensional space, thus the non-linear classification problem is well solved. SVM used kernel function to replace the inner product in original space, and mapped the data into a high-dimensional feature space. The feature space is defined by kernel function, so the construction of kernel function determines the performance of the SVM classifier [7].

At present, the kernel function used for the SVM can be divided into two categories: global kernel function and local kernel function.This paper used a combined user-defined function to improve the classification performance of SVM. Polynomial kernel function (poly) is a global kernel, good at analyzing global properties of the features; Radial Basis kernel Function (rbf) is a local kernel, taking advantages of the local properties of the features. So, taking the global and local

properties of the features into account, we construct the combination of the two functions. In this paper, the new function was combined by poly and rbf.

$$Poly : K(x, x') = ((x \cdot x') + 1)^d. \quad (4)$$

$$RBF : K(x, x') = \exp(-\|x - x'\|^2 / \sigma^2). \quad (5)$$

Mercer condition :

If $g(x) \in L_2(R^N), k(x, x') \in L_2(R^N \cdot R^N)$, $\int k \int (x, x') g(x) g(x') d_x d_{x'} \geq 0$, then $k(x, x') = (\phi(x) \cdot \phi(x'))$, so the k is the inner product of some feature space.

Construction of the kernel function that satisfied the Mercer theorem [8]:

Set both $K_1(x, x')$ and $K_2(x, x')$ the kernel function in $R_n * R_n$, then both their sum $K(x, x') = K_1(x, x') + K_2(x, x')$ and product $K(x, x') = K_1(x, x')K_2(x, x')$ are the kernel function.

By the theorem above, we can introduce:

$K(x, x') = aK_1(x, x'), K_{mix} = rK_{poly} + (1-r)K_{rbf}$ both satisfied kernel function conditions.

3.3 Optimize the Parameters

The svm toolbox used in this paper is the libsvm-mat-faruto version, in the toolbox, poly kernel has three parameters: degree (d), gamma (g), coef0 (c); RBF kernel has two parameters: c and g. d should not be too large, and typically used as 1. For a given problem, we cannot know in advance which c and g is the best; As a result, some model to choose search (parameters) is necessary.

A Genetic Algorithm (GA) is a class of adaptive stochastic optimization algorithms which involves the process of searching and optimizing. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem [9]. Genetic manipulation consists of three operators: selection, crossover and mutation.

The basic steps of GA are as follows [10]:

- 1) Under a certain encoding scheme, generate a random initial population;
 - 2) With the corresponding decoding method, convert the encoded individual into the decision variable in the problem space, and obtain the individual's fitness value;
 - 3) According to the magnitude of the individual adaptive value, select the individuals with larger adaptive value from the population to constitute the mating pool;
 - 4) Operate the individuals in mating pool through the two genetic operators crossover and mutation, and form a new generation of population;
- Repeat Step 2 ~ 4, until the convergence criterion is satisfied.

Using the heuristic algorithm GA for parameter optimization, we can find the global optimal solution without traversing all parameters within the grid, and it will take less

time. Due to the random initialization, all of the optimization results are different from each other, so several experiments should take to get the ideal c, g with better precision. There is also a selection principle that the value of c&g should be relatively small (If the c&g is too large, it can cause data overfitting, and it is not conducive to classifying the features correctly).

4 Experimental Result and Discussion

4.1 Experimental Procedure

Step1: Image acquisition. Images used in this paper are all obtained from the image library at the University of California, Berkeley.

Step2: Capture image samples. Cutting 60 small images in same size out from the original image, in which 30 images are the foreground samples, and another 30 images are the background samples.

Step3: Color feature extraction of the 60 small images. First, feature acquisition in the RGB color space and the feature vector is $T_{10} = \{R, G, B\}$, then obtain the features in the HSV color space ,combining the feature vector $T_{20} = \{H_1, S, V\}$.The feature in color space can be showed in the vector $T_1 = \{T_{10}, T_{20}\}$.

Step 4: Texture feature extraction of the 60 small images. (1)Convert each color component into grey scale. (2) In order to reduce the computation, compress the original image's grey scale to 16 levels. (3) Normalize co-occurrence matrix P as distance 1, angle of 0, 45, 90, 135 respectively. (4) Calculate the 3 texture parameters of the co-occurrence matrix: entropy, inertia moment and the correlation. (5) Calculate the mean value and the standard deviation of entropy, inertia moment and correlation as the final 6 texture feature. So, the texture feature used in this paper is $T_2 = \{I_m, H_{2m}, C_m, I_s, H_{2s}, C_s\}$.

Step 5: Using the linear function transformation, the expression is as follows:
 $y = (x - minValue) / (MaxValue - minValue)$; After the normalization, the feature vector:

$T = \{R_n, G_n, B_n, H_{1n}, S_n, V_n, I_{mn}, H_{2mn}, C_{mn}, I_{sn}, H_{2sn}, C_{sn}\}$ (R_n stands for the normalized feature R), a 1*12 vector.

Step 6: Using GA for c&g optimization. Due to the randomness of initial population, we need many times of optimization, seeking the most reasonable C and g (smaller value and the better precision).

Step 7: Convert features into SVM training data. Divide the 60 extracted features into two categories: (1) the background features(30*12 dimension), set the label as 1. (2) Set label of the 30*12 dimension foreground features as 0.

Step 8: Image segmentation based on linear combination kernel function. In this paper, 3 kernel functions are used:

$$\text{Kernel1}=0.1\text{poly}+0.9\text{rbf}; \text{Kernel2}=0.5\text{poly}+0.5\text{rbf}; \text{Kernel3}=0.9\text{poly}+0.1\text{rbf}.$$

4.2 Experiment and Discussion

(1) This section used the two images to illustrate the image segmentation result with different image features.

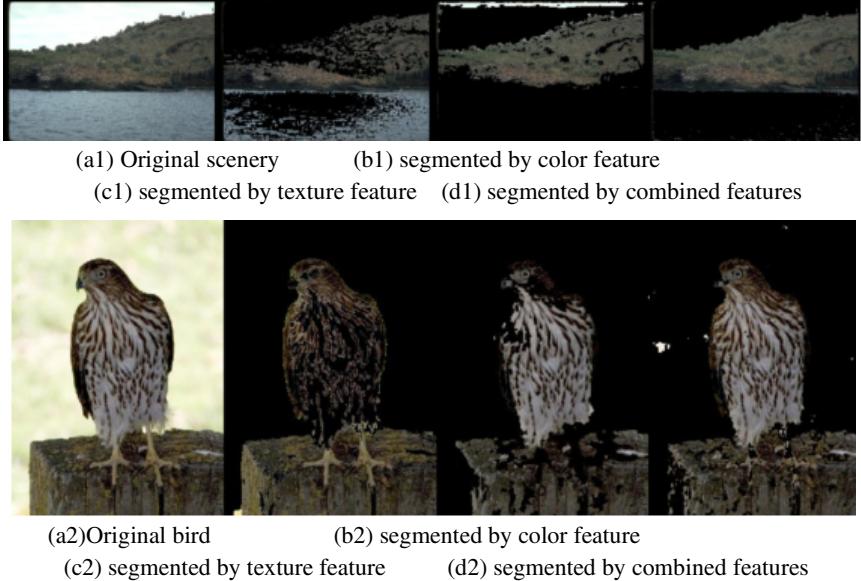
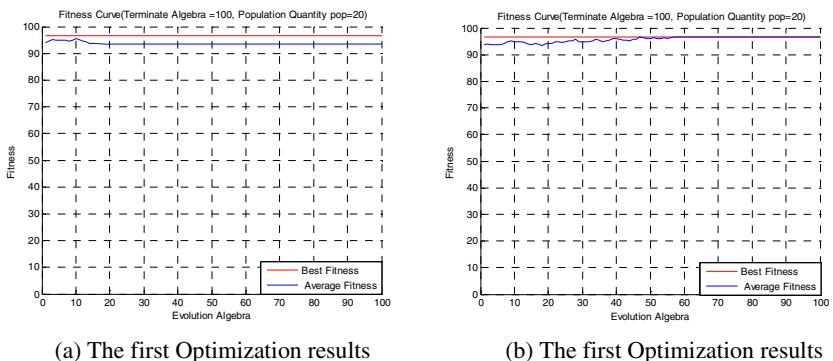


Fig. 1. Comparison of segmentation with single feature and the combined-features

It can be seen from images after segmentation, that error classification is easily caused by using a single feature, which results in poor segmentation. Therefore, using the composite features is reasonable and effective.

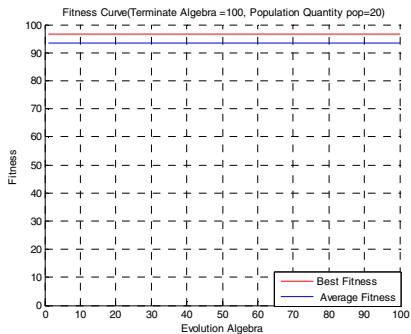
(2) Using GA for c&g optimization.



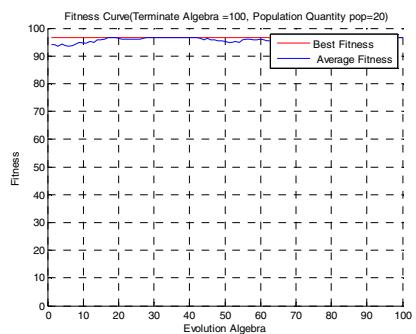
(a) The first Optimization results

(b) The first Optimization results

Fig. 2. Four optimization results of the bird. The accuracy is 96.6667%, selecting c&g with the minimum of the results. The value of c&g in the four optimization results are: 8.4482, 11.4536; 3.1672, 4.5578; 0.69149, 3.5968; 4.5575, 8.3235. We chose the third result, and the c&g are 0.69149 and 3.5968.



(c) the third Optimization results



(d) the last Optimization results

Fig. 2. (Continued.)

(3) Segmentation with different combined kernel



(a) Original bird (b) segmented by 0.1poly+0.9rbf
 (c) segmented by 0.5poly+0.5rbf (d) segmented by 0.9poly+0.1rbf

Fig. 3. The bird segmented by different combined-kernel. With the segmentation result above, we can see the picture (d) has the best segmentation result. Then we can analyze of the experimental results quantitatively with the table 1.

Table 1. The accuracy of the three combined kernels

Kernel Function	Training Accuracy	Segmentation Accuracy
0.1poly+0.9rbf	96.67% (58/60)	89.46%
0.5poly+0.5rbf	95.00% (57/60)	91.89%
0.9poly+0.1rbf	91.67% (55/60)	95.21%

We can see that when the rbf kernel has the larger weight, the training accuracy will be better and the segmentation accuracy will be worse; and the poly kernel has the contrary situation. We can combine the advantages of rbf and poly to get the combine-kernel with better segmentation performance. From the result of the segmentation, we choose the last kernel, and the accuracy is satisfied.

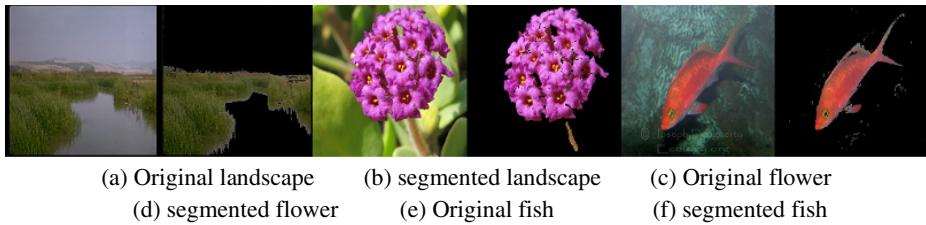


Fig. 4. Different images segmented by the method we proposed. We can see that our method is suitable for different images.

Analysis the experimental results above, when we use the full image features, the support vector machine can achieve very good training classification accuracy, effective combination of local kernel(rbf) with the global kernel(poly), we can improve the final classification results, in order to achieve a more accurate image segmentation.

References

- [1] Wang, X.Y., Wang, Q.Y.: Color Image Segmentation Using Automatic Pixel Classification with Support Vector Machine. *Neurocomputing*, 3898–3911 (2011)
- [2] Yoon, K.-J., Kwon, I.-S.: Color Image Segmentation Considering Human Sensitivity for Color Pattern Variations. In: *Proceedings of the International Society for Optical Engineering (SPIE)*, pp. 269–278 (2001)
- [3] Moreno, R., Grana, M.: Image Segmentation on Spherical Coordinate Representation of RGB Colour Space. Computational Intelligence Group, Universidad del País Vasco, IET image process 6, 1275–1283 (2012)
- [4] Zhang, J., Du., M.: Fast Image Segmentation of Gold Immunochromatographic Strip Based on Fcm Clustering Algorithm in HSV Color Space. *Image and Signal Processing(CISP)*, 525–528 (2012)
- [5] Bo, H., Ma, F.: Analysis of Calculation Problem in Image Texture Gray level Co-occurrence Matrix. *Journal of Electronics* 34(1) (2006)
- [6] Deng, N., Tian, Y.: Support Vector Machine (SVM)- Theory, Algorithms, and Development, pp. 81–86. Science Press, Beijing (2009)
- [7] Lin, H.-T., Lin, C.-J.: A Study on Sigmoid Kernels for SVM and the Training of Non-PSD Kernels by SMO-Type Methods,
<http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf>
- [8] Gao, H.Z., Wan, J.W.: The Hyperspectral Image Classification Technology Research Based on Combined Kernel in Spectral and Air Space. *Signal Processing* 27(5) (2011)
- [9] Angelina, S., Padma Suresh, L., Krishna Veni, S.H.: Image Segmentation Based On Genetic Algorithm for Region Growth and Region Merging. In: *2012 International Conference on Computing, Electronics and Electrical Technologies*, pp. 970–974 (2012)
- [10] Liu, G.H., Bao, H., Li, W.C.: Using Matlab to Realize Genetic Algorithm Program. *Computer Application and Research*, 80–82 (2001)

Co-expressing Patterns of Schizophrenia Candidate Genes in Brain Regions

Xinguo Lu^{1,2}, Bingtao Feng¹, Yong Deng¹, and Dewen Hu²

¹ School of Information Science and Engineering, Hunan University, Changsha, 410082, China

² College of Mechatronics and Automation,
National University of Defense Technology, Changsha, 410073, China
hnluxinguo@126.com

Abstract. Transcriptional profiling of human brain will be great helpful understanding the molecular mechanism of schizophrenia, which is a severe threaten to human health. In this study, top 250 schizophrenia candidate genes were selected, and then expression pattern of these 250 schizophrenia candidate genes were analyzed in six brain regions, including frontal lobe, temporal lobe, parietal lobe, basal ganglia, occipital lobe, and hippocampus by 2D hierarchical clustering method. We also used stability analyzing to evaluate clustering methods, and pearson-pairwise correlation analysis was executed between each two different tissues in each sub-region. 2D hierarchical clustering results indicated certain gene pathology have similar expression level in some brain sub-regions while different in some others. We found that local pattern of the sample clustering can reflects the cytoarchitecture of basal ganglia, hippocampus and occipital lobe, while in temporal lobe, parietal lobe and frontal lobe the situation is less discriminable. As respect to the gene cluster pattern, we found from the GO term of each gene cluster that there are some genes have the similar co-expressing level across some brain regions, while in some other regions, the results we found is just the opposite that the co-expressing level appears to be very different. Our experimental results strongly proved that local pattern of schizophrenia candidate genes were not only just simply reflecting sub-brain cytoarchitecture but also on functional coordination between each molecular components. And those co-expression pattern in all the brain regions, it may be helpful to understand the pathology mechanisms of schizophrenia related to each sub-brain.

1 Introduction

Schizophrenia have been a very important threaten to human health. It is a key step to elucidate the enormous complexity of the human brain which is a function of its precise circuitry, its structural and cellular diversity, and, ultimately, the regulation of its underlying transcriptome. Some studies of the developing human brain have used relatively small numbers of samples and predominantly focused on only a few regions or developmental time points. Kang reported the generation

and analysis of exon-level transcriptome and associated genotyping data, representing males and females of different ethnicities, from multiple brain regions and neocortical areas of developing and adult post-mortem human brains[1]. Recently, Hawrylycz describe the generation and analysis of a transcriptional atlas of the adult human brain, comprising extensive histological analysis and comprehensive microarray profiling of 900 neuroanatomically precise subdivisions in two individuals[2].However, there are limited studies on transcriptional profiling of schizophrenia from multiple brain regions and neocortical areas. Allen carried out a systematically random-effects meta-analyses and have created a regularly updated online database of all published genetic association studies for schizophrenia[3].In the present study, we selected top 250 schizophrenia candidate genes from the online database SzGene, and then collected form transcriptional profiling data of six brain regions from Allen Human Brain Atlas¹. the results of hierarchical 2D cluster, pearson correlation and connectivity test among the six regions will have great benefits for understandings of molecular mechanism of schizophrenia.We found some significant co-expressing phenomena in both gene cluster pattern and sample cluster pattern of schizophrenic candidate genes expression data in six regions,including basal ganglia the so called expression center of schizophrenia candidate genes[4].

2 Materials and Methods

2.1 Data Collection

All the data in this study were collected from Allen Human Brain Atlas ². In brief,The raw data were generated Postmortem brain from males and females between 18 and 68 years of age, with no known neuropsychiatric or neuropathological history. After Initial collection, dissection and freezing of whole brain, Large Format Sectioning and Blockface Imaging, Slab Partitioning, the data were generated by using an Agilent 8x60K array, custom-designed by Beckman Coulter Genomics in conjunction with the Allen Institute.

2.2 Top 250 Schizophrenia Candidate Genes

The SZGene database⁵ contains information from 1727 studies, reporting data on 1008 genes, and 8788 polymorphisms; this database has 287 meta-analyses.SZGene ranks its top results using the HuGENet interim guidelines published by Ioannidis and colleagues, which consider the amount of evidence[3][7]. We selected top 250 schizophrenia candidate genes from above dataset and used them in the following work.

¹ <http://www.szgene.org/>

² <http://www.brain-map.org/>

2.3 Data Analysis

We executed unsuperized hieratical 2D cluster on the purpose dataset.of the top 250 schizophrenia candidate genes, and then calculated correlation matrices of pairwise comparisons among regions,this analysis explained the undiscriminility of clustering.

2.4 Validation of Hieratical 2D Cluster

We evaluate the cluster by meassure it connectivty which should be minimized[8]. the result of our validation proved that the hieratical cluster is the most suitable algorithm for our dataset.The connectivity has a value between zero and one and should be minimized.

2.5 Gene Ontology Analysis

We discorvered from the 2D hieratical clustering results that the gene expressing clust parttern show significantly co-expressing in every single sub-brain region we tested.GO analysis has been executed on every bunch of clustered genes.To test these genes for over-representation of GO terms,we computes for all GO nodes a hypergeometric distribution test and returns the corresponding raw and Bonferroni corrected p-values. A subsequent filter function performs a GO Slim analysis using default or custom GO Slim categories[9]. We provides similar utilites as the hyperGTest function in the GOstats package from BioConductor³.The main difference is that Our Method simplifies the usage of custom chip-to-gene and gene-to-GO mappings.Every bunch of clustered genes we obtained 6 GO terms,we choose the most Enriched GO term according the p-value and put it into our results.

3 Results

2D hierarchical clustering of brain tissue and schizophrenia top 250 gene is conducted for six brain regions: temporal lobe, basal ganglia, hippocamel, occipital lobe, parietal lobe and frontal lobe, respectively. Pearson correlation coefficients of each region are calculated one by one. A connectivety test is performed for application effect of experimental data of 2D hierarchical clustering, which are compared with two clustering methods of pam and kmeans.

3.1 2D Hierarchical Clustering Results of Basal Ganglia

Basal ganglia 2D hierarchical clustering results indicate that compared with CPi and CPe, cl has good discrimination, but CPi and CPe are discriminated

³ <http://www.bioconductor.org/>

quite not obviously. This is the justification of their similar molecular organization. The clustering results reflect that some genes, interestingly and manifestly, present coordinate expression. GO term enrichment analysis results are shown as follows: 1. interleukin-3 receptor binding; 2. insulin receptor binding; 3. interleukin-4 receptor activity; 4. protein-L-isoaspartate (D-aspartate) O-methyltransferase activity; 5. beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1, 6-N-acetylglucosaminyltransferase activity (Fig. 1a). Euclidean distance of log2-transformed signal intensity was used to measure pairwise similarity. The calculated results of pairwise pearson correlation coefficient of three kinds of component genes indicate that there is comparatively high correlation between CPi and CPe (Fig. 1b). As shown from the results of linear graphs of Connectivity index calculation of c, clustering effectiveness evaluation, three clustering methods, kmeans, pam and hierarchical, via horizontal comparison, are most appropriate for the data set, and they begin to converge as the clustering number is greater than 15 (Fig. 1c).

3.2 Results of the Rest Five Brain Regions

We obtain the results of the same multiterms of the Hippocampel,Occipital lobe,Temporal,Parietal lobe,Frontal lobe which had been list in table1 and table2. As about to the clustering effectiveness evaluation, Among three clustering methods, kmeans, pam and hierarchical, via horizontal comparison,hierarchical are most appropriate for the data set for all these regions we choose.

Table 1. Discriminability of the local Pattern($P=0.05$)

brain region	discriminability	correlation
Basal ganglia	high	low
Hippocampal	high	low
Occipital lobe	high	low
Temporal lobe	low	high
Parietal lobe	low	high
Frontal lobe	low	high

4 Discussion

The reason why we select 250 schizophrenia genes is trying to explore these genes role in those brain regions. The finally clustered gene clusters were treated with GO pathway enrichment for analyzability of results. We modularized the clustering results of each brain regions tissue by divide-and-conquer method. Through the WGCNA analysis for each module, we obtained the correlation value[4][9][10]. Hyo Jung Kang and others used pairwise Spearman correlations between brain regions (top) and between NCX areas(bottom) during fetal development (periods 3C7), postnatal development (periods 8C12) and adulthood

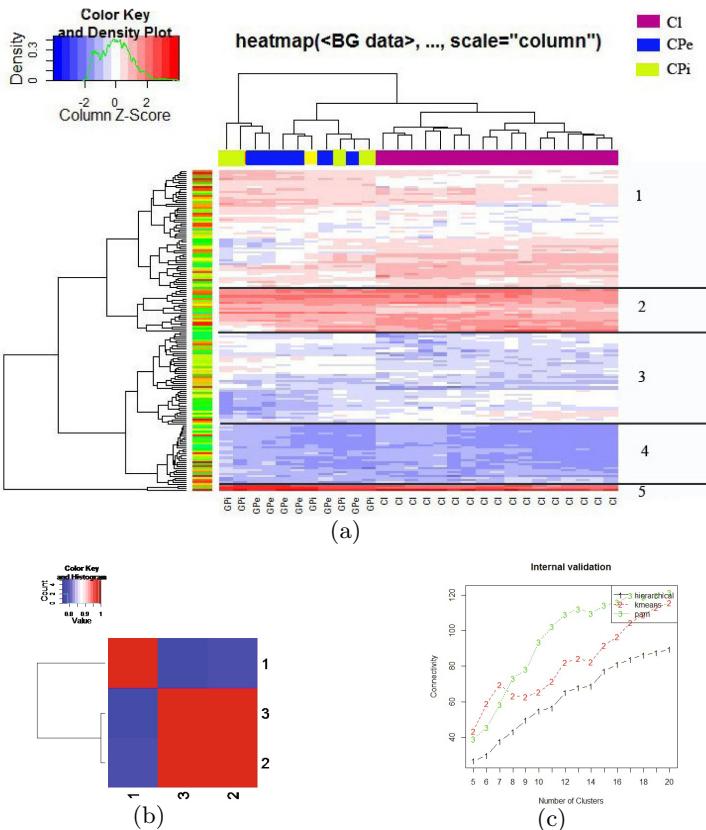


Fig. 1. 2D hierarchical clustering and identification of basal ganglia 2D hierarchical clustering results pearson correlation coefficient,1:Cl,2:CPi,3:CPe.c,Connectivity test

(periods 13C15) to calculate the pearson pairwise correlation coefficients of gene expression between each brain region in each development stage. We, however, calculated the pearson pairwise correlation coefficients between different molecular organizations in six brain regions and the results has a good corresponding relation with the clustering results of molecular organization. Michael using 2D hierarchical clustering, proved that the molecular organization in the hippocampal area correspond to the local clustering model[2]. In the test that we conducted, 2D hierarchical clustering is applied in six brain regions, including the hippocampal area, and we further found that the local clustering model of the hippocampal area, basal ganglia and occipital lobe has a manifest corresponding relation with the molecular organization. The other three regions, however, are in a mixed state and present not so obviously. We used the inline of cluster as the clustering quality test to prove that the hierarchical clustering is most appropriate for the data set[8].

Table 2. The degree of expression of GO terms among brain regions

GO term	Hippocampal	Occipital lobe	Temporal Parietal lobe Frontal lobe		
interleukin-3 receptor binding	high	high	medium	high	high
interleukin-4 receptor activity	null	medium	high	low	null
protein-L-isoaspartate (D-aspartate)	high	null	low	null	null
beta-1,3-galactosyl-O-glycosyl-glycoproteinl	null	null	low	null	null
ribosomal large subunit binding	low	low	high	low	low
high-density lipoprotein particle receptor binding	low	null	null	high	high
myristoyltransferaseactivity	high	low	high	medium	high
squalene synthase activity	null	null	null	null	medium

Some researchers found that genes of specific expression mainly concentrate in the Basal ganglia and accordingly considered that basal ganglia would be the expression center of schizophrenia candidate genes and provided the intrinsic pathophysiology of schizophrenia[8]. We believed that basal ganglia, hippocamel and occipital lobe have manifest division of function of each molecular organization. The other three regions, temporal lobe, parietal lobe and frontal lobe, however, have mixed clustering results. Then we calculated the pearson correlation coefficients one by one in each region. As presented from above research results, in the basal ganglia, frontal lobe, hippocamel and parietal lobe, Interleukin-3 receptor binding has higher coordinate expression and has much higher coordinate expression in occipital lobe, while has lower coordinate expression in temporal lobe. Ribosomal large subunit binding GO term has medium coordinate expression in two regions: hippocamel and parietal Lobe and has higher coordinate expression in temporal lobe and frontal lobe, and lower coordinate expression in occipital lobe. High-density lipoprotein particle receptor binding has medium coordinate expression in frontal lobe and hippocamel and lower in parietal lobe. Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyl transferase activity has higher coordinate expression in basal ganglia, very low in two regions of temporal lobe. Interleukin-4 receptor activity has medium coordinate expression in parietal lobe and occipital lobe and higher in basal ganglia and temporal lobe. Protein-L-isoaspartate (D-aspartate) O-methyltransferase activity has lower coordinate expression in basal ganglia and temporal lobe. Myristoyltransferase activity has higher coordinate expression in hippocamel, parietal lobe and temporal lobe and lower in occipital lobe. With this discovery may be we can speculate that different sub-brain region have the same or at least similar fuctions which some how affect the schizophrenia, or just the opposite.

5 Conclusion

In the present study, we analyzed the human gene expression levels among frontal lobe, occipital lobe, parietal lobe, temporal lobe, basal ganglia and hippocamel and explored the reflection between the local sample cluster pattern of schizophrenia candidate genes of expression data and the cytoarchitecture of human brain regions. Furthermore, we discovered, which is obvious, that there

is some co-expression phenomena of the local gene cluster pattern as well. we found that in basal ganglia,hippocamel and occipital lobe the reflection between local pattern and cytoarchitecture is clear and discriminable,while in temporal lobe,parieatal lobe and frontal lobe this reflection is gong fuzzy.All this results are perfectly coordinate with the pearson correlation analysis output we carried on each brain region.According the co-expressing local gene cluster we made GO-analysis on them and found some GO terms are enriched to among different brain regions,some them performance high expressing level across some brain regions and in other brain regions are low.We therefore conclude that local pattern of schizophrenia candidate genes were not only just simply reflecting sub-brain cytoarchitecture but also on functional coordination between each molecular components.And those co-expression pattern in all the brain regions,it may be helpful to understand the pathology mechanisms of schizophrenia related to each sub-brain.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (61202288), the Ph.D. Programs Foundation of Ministry of Education of China (20100161120023), the National Science Foundation for Post-doctoral Scientists of China (20100471790), the Fundamental Research Funds for the Central Universities and the Young Teachers Program of Hunan University.

References

1. Kang, H.J., Kawasawa, Y.I., Feng, C., et al.: Spatio-temporal transcriptome of the human brain. *Nature* 478, 483–489
2. Hawrylycz, M.J., Lein, E.S., Guillozet-Bongaarts, A.L., et al.: An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489, 391–399 (2012)
3. Allen, N.C., Bagade, S., McQueen, M.B., et al.: Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the szgene database. *Nat. Genet.* 40(7), 827–834 (2008)
4. Lu, X., Liu, P., Zeng, L.-l., Li, R., Hu, D.: Schizophrenia Candidate Genes Specific to Human Brain Region are Restricted to Basal Ganglia. In: Yang, J., Fang, F., Sun, C. (eds.) IScIDE 2012. LNCS, vol. 7751, pp. 565–572. Springer, Heidelberg (2013)
5. Sharma, A., Imoto, S., Miyano, S., et al.: A Top-r Feature Selection Algorithm for Microarray Gene Expression Data. *IEEE/ACM Transactions on computational biology and bioinformatics* 9(3), 754–764 (2012)
6. Melendez, R.I., McGinty, J.F., Kalivas, P.W., et al.: Brain region-specific gene expression changes after chronic intermittent ethanol exposure and early withdrawal in C57BL/6J mice. *Addict. Biol.* 17(2), 351–364 (2012)
7. Bilder, R.M., Howe, A., Novak, N., et al.: The genetics of cognitive impairment in schizophrenia: a phenomic perspective. *Trends Cogn. Sci.* 15(9), 428–435 (2011)
8. Brock, G., Pihur, V., Datta, S., et al.: clValid: An R package for cluster validation. *Journal of Statistical Software* 25(4) (March 2008)

9. Horan, K., Jang, C., Bailey-Serres, J., et al.: Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.* 147(1), 41–57 (2008)
10. Zhang, B., Horvath, S., Langfield, P., et al.: General framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4 (2005)
11. Horvath, S., Zhang, B., Carlson, M., et al.: Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* 103, 17402–17407 (2006)

Boosting Deformable Part Model by Sample Sharing and Outlier Ablation

Feng Liu¹, Yongzhen Huang², Liang Wang², and Wankou Yang¹

¹ School of Automation, Southeast University, Nanjing, 210096, China

² National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

Abstract. The deformable part model (DPM) achieves the best performance on some well known datasets in terms of object detection. Literature springs up to study the success of such a model and hence various methods are proposed to improve it. Yet one import issue, the sensitivity to outliers of the hinge loss,¹ has not been fully studied. In this paper, we take two initiatives to handle this problem: 1) we propose to share samples of one component to others by similarity; 2) we give samples different weights according to their costs. The model is better trained with our proposed method, and we boost the performance of the newly released voc-release 5 [6] model on the challenging VOC 2007 dataset.

Keywords: Object Detection, Deformable Part Model, Sample Sharing, Outlier Ablation.

1 Introduction

Object detection aims to find the object bounding boxes of the target class in a picture, and the deformable part model [3] is one of the most well-known models in this field. The DPM consists of several components with each component be a star model which is made of a root filter and several part filters. Usually, each component has a different aspect ratio² in order to catch objects of changing postures, viewpoints and deformations.

Divvala et al. [2] have made a comprehensive evaluation of the DPM and claim that, rather than the usage of deformation parts, the utilization of multiple components seems to contribute the most to achieve a high performance. The reason is that a single linear hyperplane is not capable of separating the sophisticated objects from the backgrounds, as the positive samples tend to scatter in the feature space. Using multiple components is to separate the feature space with several hyperplanes, each of which in charge of a subset of samples with some common properties. For example, in the DPM, samples of the same component have similar aspect ratios, and in Divvala's work, such samples are close in the feature space by Euclidian distance. That is to say, it is better for the samples

¹ DPM is typically trained by the latent SVM, the loss of which is hinge loss.

² $aspect\ ratio = height / width$

to have small variations in the same component. Zhu et al. [12] also demonstrate that ‘clean’³ data can help to improve the performance of a mixture model, and the ‘clean’ data is obtained by clustering. This viewpoint is also supported by the poor performed classes of the DPM on the VOC 2007 dataset, e.g., cat. These classes tend to have more deformations, which lead to a large variation in each component, thus causing an ill-trained model.

Why does large intra-component variation deteriorate the performance, especially for the SVM? It is because hinge loss is more sensitive to outliers compared with the 0-1 loss [11]. The loss suffered is proportional to the distance of a point to the classification hyperplane. In this condition, the hyperplane is more likely to turn towards the outliers to compensate the loss caused by them. In this paper, we consider samples which are dissimilar with the majority of the samples of this component as outliers, and take two initiatives to alleviate the effects of intra-component impurity caused by them:

1. Sample sharing: Instead of restricting one positive sample to be assigned exactly to one component, we relax this constraint by allowing a sample to be used by multiple components according to its similarity with them. Thus each component has more samples to choose.
2. Outlier ablation: A positive sample is given a different weight for each component. The weights of all samples as well as the SVM parameters are jointly learned by minimizing an augmented loss function. In this way, the samples which cause large losses will be given small weights (or be abandoned). The model can thus gain robustness to outliers.

The rest of the paper is organized as follows: we firstly revisit the related work in section 2. Then we propose our method in section 3. Our experimental results are reported in section 4 before we draw a conclusion in section 5.

2 Related Work

The deformable part model [3] is the state of art object detector on many datasets. Papers are published to study its success and lots of improvements are made. Among them, some [7,10] propose to combine the DPM with other cues to boost the overall detection results. Some [1,2] propose to improve the model with carefully initialized components or a learned parts relationship by using finer annotations.

Specially, Divvala et al. [5] attribute the DPM’s success largely to the usage of multiple components. They split components by clustering the HOG features of all samples, replacing the aspect ratio heuristic used by the original DPM. With a more reasonable component initialization, the samples assigned to each component become ‘cleaner’, thus the model gets better trained. Zhu et al. [12] evaluate the influence of a variety of factors to the object detector’s performance and advocate that we need ‘clean’ data to train our model. They also show that

³ The data is called ‘clean’ when they have small variations.

clustering all the samples (by K-means) and using samples of the same cluster to train a separate component will help. They ascribe the bad performance of the model trained using unpurified data to the hinge loss, which is vulnerable to outliers. Our approach differs from theirs because we try to start directly from the defects of the hinge loss by modifying the loss function and generate virtual data by sample sharing. So our model does not depend on specific component initialization strategies. Instead, our method is easy to combine with them.

Our work is also related with [8] which tries to share samples between datasets and [9] which shows the possibility of training an object model with only one positive sample (yet many negative samples). However, both the final goal and the means between their methods and ours are different.

3 The Proposed Method

In spite of great achievements the DPM has made, it still performs not so well on some classes whose instances vary greatly in appearance, e.g., cat, cow. Keeping the same number of components will make these classes have a larger intra-class variation than others. However, it will be more complex and slow to inference if we use a model with more components. Moreover, the total number of instances hinders us from doing this. So we seek to improve the model without changing the number of components used. Fortunately, we find in [12] that the sensitivity of hinge loss might partly responsible for the degraded performance by impure data. We also find that better trained root filter will benefit the part initialization process, which is crucial to achieve a good performance by latent SVM. Before proposing the improving strategies, we firstly review the training process of the DPM.

3.1 Deformable Part Model Revisited

The Deformable part model is composed of several components where each component is a star model which in charge of a specific subset of objects for a class. Typically, one component includes a root filter and multiple part filters. During training only the bounding boxes of an object is observed, so the component index and part location must be inferred by the model. Thus the model is trained by the latent SVM in a stagewise way which behaves like the EM algorithm.

Stage 1. The positive samples (objects annotated by bounding boxes) are split into m groups according to their aspect ratios. For each group of positive samples, we train a separate SVM with randomly chosen negative samples. The learned weights are used to initialize the root filters of the mixture model.

Stage 2. The mixture model with only root filters is refined by letting each component choosing positive samples which fit it best. In this step, each positive sample can only be and must be assigned to one component for retraining.

Stage 3. The deformation parts are initialized from the refined root filter by an energy coverage rule. Then the new model is retrained on the full dataset with latent detection and some data-mining techniques for the negative samples.

3.2 Sample Sharing

In object detection, each object is annotated with a bounding box in an image, and we call it an object level sample x . In general, x can be used by only one component M_c , and we denote the sample for the c -th component as x^c . However, the strategies of assigning a sample to a specific component is usually heuristic or not precise, e.g., aspect ratio, especially in the first stage. What's more, for some object classes, samples of different components are similar after an appropriate transform, e.g., resizing a bus from this aspect ratio to another. If a sample fits several component models well and we just use it to train one of them, it is a waste of samples. So we draw samples for all component models from a single bounding box and this is called *sample sharing*.

The benefit of sample sharing is that we can increase the number of samples a component can choose, and it is implemented as follows. For the first stage, we just resize the sample of this component to another, since all positive samples are warped. As to the stages afterward, we allow each component to extract a sample x^c from a particular bounding box x . Note that to be selected, x^c has to be the highest scoring hypothesis for this component and satisfies the overlapping rule. The usage of a sample is determined by the strategy illustrated in the next subsection.

3.3 Outlier Ablation

For a class of large intra-class variation, it is difficult for a model M_c to account for all positive samples assigned to it, because hinge loss is rather sensitive to outliers.⁴ Putting too much efforts to fit points away from the majority of the data points will deflect the classification hyperplane to the outliers, thus more low cost data points are misclassified. Though the total loss decreases, the error rate, however, increases. Intuitively, those points causing great losses should be abandoned or be given small weights to prevent them from dominating the total loss. So instead of trying to fit each positive sample well, we might as well focus on those representative ones. We now select samples from the candidate samples set by optimizing the following objective function:

$$\min_{\beta, \alpha} \sum_{c=1}^m \sum_{i=1}^{n_p} \alpha_{c,i} \ell(\beta_c, x_i^c, +1) + \sum_{c=1}^m \sum_{j=1}^{n_n} \ell(\beta_c, x_{j,c}, -1) + \lambda_1 \|\beta_c\|_2^2 + \lambda_2^c \mathcal{R}(1-\alpha), \quad (1)$$

where $\alpha_{c,i} \in [0, 1]$ is the weight of x_i^c ,⁵ and it is initialized to one if x_i originally belongs to component c , otherwise it is set to zero. β_c is the parameter for the c -th component, and m, n_p, n_n are the number of components, positive instances and negative samples respectively. $\mathcal{R}(\cdot)$ is a regularization term which can either be the l_1 norm or the l_2 norm. $\ell = \max(1 - y_i \beta_c^T x_i^c, 0)$ is the hinge loss. λ_1, λ_2^c are hyperparameters which control the weights of the regularization terms.

⁴ In this paper an outlier is a data point which is far away from the majority of the data points, not limited to those mislabeled data.

⁵ x_i^c is the shared sample of x_i for component c and x_i is the i -th object sample. $x_{j,c}$ is the j -th negative sample for component c .

In this way, a positive sample will be given a small weight if its cost is very large. The hyperparameter λ_2^c can control the total members of samples used for each component, which means the bigger the λ_2^c , the larger amount of samples will be chosen. Compared with a standard SVM, the slope of the above objective is much smaller when the weight multiplied is less than one, which makes the model robust to outliers. One can verify the above objective is an upper bound of the 0-1 loss when l_1 norm is used and λ_2 is larger than one. So we opt to take l_1 norm as our regularizer.

The objective is biconvex, which is convex when optimizing one parameter while fixing the other. Suppose we have trained β and obtained the cost for each sample. Let us denote the cost as $L = (\ell_1^1, \dots, \ell_1^m, \dots, \ell_n^1, \dots, \ell_n^m)$, then the problem becomes: $\min_{\alpha} L^T \alpha + \sum_{c=1}^m \lambda_2^c (1 - \alpha^c)$ which is a linear programming problem, and we can get a analytical solution by coordinate descent. The answer is : $\alpha_i^c = 0$, if $\ell_i^c > \lambda_2^c$ and $\alpha_i^c = 1$, otherwise. In practice we take the top scoring p percentage samples to train one component and regard the rest as outliers. The theoretical support for being able to adopt this alternative measure is that for the positive sample, the lower the score the higher the loss. If the loss of a sample x_i^c is above λ_2^c , it will be dropped. It is equal to take the top scoring ones. This is how *outlier ablation* takes place. Similar measures have also been taken by Gaidonet al. [4], however, no theoretical analysis is given in his work.

3.4 DPM Equipped with the Proposed Strategies

We now integrate the two strategies developed above into the DPM. Specifically, we employ both methods on the first stage. We firstly initialize the root filter according to the aspect ratio heuristic and score the samples of the candidate samples set by the trained filters. The top scoring $p_1\%$ samples are chosen to retrained the model. On the second and third stage, we just use the outlier ablation strategy, because in practice, we find that the shared samples always score lower than the original ones. The percentage of samples used in stage two and stage three are denoted as $p_2\%$ and $p_3\%$ respectively. Usually, we take $p_1 < p_2 < p_3$.

4 Experiments

In this section, we firstly give an introduction of the dataset and experimental settings in 4.1, then we show some primacy results on the cat class of the VOC 2007 dataset⁶ in 4.2. At last, the results of all classes are reported in 4.3.

4.1 Dataset and Experimental Settings

The PASCAL VOC 2007 dataset is chosen to evaluate our proposed model. It is a challenging dataset which consists of thousands of images of real world scenes

⁶ “<http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>”

Table 1. Performance achieved by the 5 models on the VOC 2007 cat class

models	M1	M2	M3	M4	M5
AP	11.8	14.2	23.0	24.9	25.5

over 20 classes. The unbalanced distribution of images over classes and large intra-class variations both make it a tough task.

Our model is composed of 3 components with each component of 8 parts. In the first stage, the number of samples of a component c to retrain the model is set to $0.3n_c$, where n_c is the amount of samples assigned to this component initially. We retrain the model of the first stage twice. The retraining of the second stage loops for two times with $0.6n_c$ samples. For the third stage, we firstly use $0.6n_c$ samples for six rounds and use all samples in the next three rounds. The purpose is to firstly train the appearance filters well with ‘clean’ data and then use all the samples to learn the parameters of the deformation features. The hyperparameters p_1, p_2, p_3 are tuned on the validation set. Though setting different values of these parameters for each class may get better performances, we use a unified term of value to ensure the fairness of comparison.

4.2 Some Primary Results

Before testing it on all classes, we firstly test several models on the cat class of the VOC 2007 dataset. The reason for using this class is that cats are flexible objects and this class is of large intra-class variation, thus it is perfect for testing our proposed method. We compare the performance of the following methods:

M1: A mixture model with 3 components, and each component just contains a root filter. It is trained using the framework of voc-release 5 [6], however, no parts are added. This corresponds to training a DPM which just uses the first two stages.

M2: A mixture model similar to M1. However, we use the proposed two strategies in the first stage and outlier ablation in the second stage.

M3: The original DPM implemented in voc-release 5 which contains 3 components with each component be of 8 parts.

M4: The DPM following the same settings as M3, yet with the proposed two methods used in the first stage and outlier ablation used in the second stage.

M5: The DPM as described in section 4.1.

We list the average precision (AP) achieved by each model in Table 1, and visualize the models trained by the first two methods in Fig. 1. From the table, we can see that our method significantly improve the mixture model, especially the one without parts. Possible reason is that a simple model doesn’t have the ability to handle large variations, so the model misclassifies many easy samples in order to catch the outliers. Our model just focuses on the representative samples,

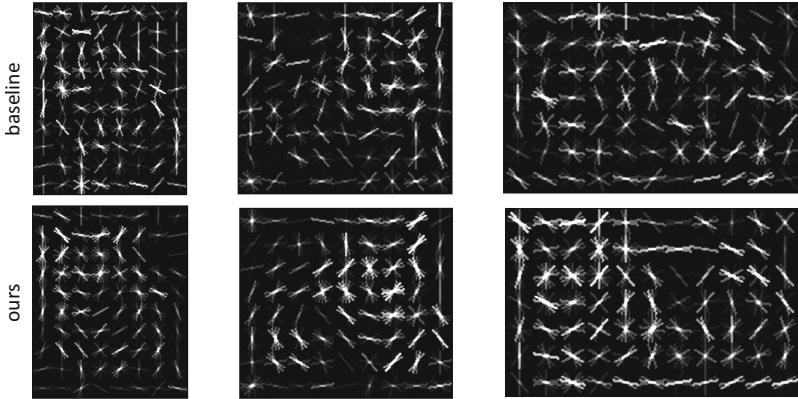


Fig. 1. Visualization of the original mixture model (M1, top row) and the one adopted our improvements (M2, bottom row) on the cat class. Our model has a more clear contour and has large weights on some specific locations. One can see that each component corresponds to a particular posture.

Table 2. Comparison of the original DPM with the improved model on the full VOC 2007 dataset (better performing ones are in bold). Our model beat the original DPM on most classes, especially those of large intra-class variation.

class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
M3 [6]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7
M5	33.1	59.0	12.5	17.5	26.2	55.3	57.7	25.5	21.7	27.1	32.6
<hr/>											
class	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean	
M3 [6]	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7	
M5	13.2	59.7	46.5	42.3	14.0	23.8	35.2	44.3	39.0	34.3	

so the performance becomes better. After adding parts, a model can handle more variations, so the margin becomes smaller. The boosted performance of M4 over M3 proves that the initialization of part filters given by our model is better than the original one. With a good initialization, the latent SVM gets better trained. This can also be illustrated in Fig. 1 as our model learns a much clearer contour and is quite confident about specific gradient orientations at a fixed location.

4.3 Results on the PASCAL VOC 2007 Dataset

To show the effectiveness of our proposed method, we evaluate it on all classes of the VOC 2007 dataset, and the results are reported in Table 2. From the table we can see that our method outperforms the original DPM on 11 classes,

and the margin is quite large on classes which have large intra-class variations. However, there still exist some classes on which we fail to beat the original model. The reason is that training a model is a tradeoff between data purity and the number of training samples [12], the performance will decrease if we use only a subset of the ‘clean’ data. Consequently, the mean average precision over all classes improves by 0.6. This improvement is impressive since the newly released voc-release 5 is the best object detector when only HOG feature are used.

5 Conclusion

In this paper, we give a theoretical analysis on how data impurity affects the DPM and propose two strategies, sample sharing and outlier ablation, to alleviate the harms caused by it. The sample sharing strategy seeks to enlarge the candidate samples set size for each component, while the outlier ablation scheme tries to fit only the representative samples by dropping the outliers. With the above strategies, models of most classes get better trained, especially those with large intra-class variation. Currently, however, there are no effective ways to choose the hyperparameters for sample selection, and we can only determine them by validation. So in future, we want to develop an explicit purity measure and find a way to automatically learn the hyperparameters.

Acknowledgement. This work is jointly supported by National Natural Science Foundation of China (61175003, 61135002, 61203252, 61005008, 61273023), Hundred Talents Program of CAS, and Tsinghua National Laboratory for Information Science and Technology Cross-discipline Foundation.

References

1. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*, Part I. LNCS, vol. 7572, pp. 836–849. Springer, Heidelberg (2012)
2. Divvala, S.K., Efros, A.A., Hebert, M.: How important are “Deformable parts” in the deformable parts model? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) *ECCV 2012 Ws/Demos*, Part III. LNCS, vol. 7585, pp. 31–40. Springer, Heidelberg (2012)
3. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *TPAMI* 32(9), 1627–1645
4. Gaidon, A., Marszalek, M., Schmid, C., et al.: Mining visual actions from movies. In: *BMVC* 2009 (2009)
5. Gao, T., Stark, M., Koller, D.: What makes a good detector? – structured priors for learning from few examples. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*, Part V. LNCS, vol. 7576, pp. 354–367. Springer, Heidelberg (2012)
6. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5,
<http://people.cs.uchicago.edu/~rbg/latent-release5/>

7. Gu, C., Arbeláez, P., Lin, Y., Yu, K., Malik, J.: Multi-component models for object detection. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 445–458. Springer, Heidelberg (2012)
8. Lim, J.J., Salakhutdinov, R., Torralba, A.: Transfer learning by borrowing examples for multiclass object detection. In: NIPS 2011 (2011)
9. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV 2011, pp. 89–96 (2011)
10. Mottaghi, R.: Augmenting deformable part models with irregular-shaped object patches. In: CVPR 2012, pp. 3116–3123 (2012)
11. Xu, L., Crammer, K., Schuurmans, D.: Robust support vector machine training via convex outlier ablation. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21, p. 536
12. Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C.: Do we need more training data or better models for object detection? In: BMVC 2012 (2012)

Multi-cue Visual Tracking Based on Sparse Representation

Xiping Duan^{1,2,*}, Jiafeng Liu¹, and XiangLong Tang¹

¹ School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, China

² College of Computer Science and Information Engineering, Harbin Normal University,
Harbin 150025, China

xpduan_1999@126.com, {jefferyliu, tangxl}@hit.edu.cn

Abstract. Under dynamic and complex environment, the single feature methods usually can't distinguish the target from background well, so that multiple features are considered in the paper. For each candidate, multiple features are extracted and conducted the sparse representation respectively, then observation probability is calculated by combinatting reconstruction errors of multiple features in particle filter framework. Comparing with single feature method, the proposed method performed robust with better accuracy. And further experiments on some representative image sequences showed that the proposed method also performs well in complex scenarios, such as varying illumination, background clutter, and occlusion.

Keywords: Computer vision, visual tracking, particle filter, sparse representation.

1 Introduction

Tracking a target in a video, as a hot topic in computer vision, arises in many applications such as automatic surveillance, vehicle navigation, advanced human-computer interaction and many others. For the past several decades, a variety of related tracking algorithms appeared. The details can be referred to in [1]. Due to the intrinsic factors such as pose variation and shape deformation, and extrinsic factors such as background clutter, varying illumination, noise, occlusion and so forth, the robust target appearance representation is critical and desired.

Recently, sparse representation has played an important role in face recognition, background subtraction, texture segmentation and so forth, and has been extended to the visual tracking. In visual tracking, the target candidate can be represented as a sparse linear combination of dictionary templates, and its reconstruction error, as the similarity measure with target, is used in computing the observation probability. The appearance representation can be embedded into a tracking framework to achieve tracking a target in video sequence. The previous related works[2][3][4][5] of sparse

* Corresponding author.

representation based tracking, is robust to noise, occlusion and other image corruptions. However, these trackers need to solve computationally expensive L_1 norm regularized least squares problem for every target candidate of every frame, so computational burden become its bottleneck in many application areas. Another drawback is that most of these trackers are based on single feature such as image intensity and ignore other features, which is prone to causing tracking drift potentially in some complex scenarios.

Based on the above analysis, a new tracking method, based on multi-cue sparse representation, is proposed. This method extracts multiple features for each candidate, and then conducts the sparse representation respectively for each feature. Different from the previous works based on single feature sparse representation, the proposed method not only incorporates sparse representation, but also utilizes the rich multiple features. Thus, the proposed method can achieve mutual complementation among multiple features.

2 Multi-cue Sparse Representation

2.1 Notation and Target Appearance Representation

Let $T = \{T_1, T_2, \dots, T_J\}$ denotes the target template set, composed of J previous tracking targets. K features extracted from each target template, among which $T^k = \{T_1^k, T_2^k, \dots, T_J^k\}$, $k=1, 2, \dots, K$, denotes the k th template set corresponding to the k th feature. T_j^k , the j th entry of T^k , denotes its j th template, which corresponds to the k th feature of the j th target template T_j . In this case, each feature y^k , $k=1, 2, \dots, K$, of each sampled candidate y can be represented as the combination of templates of the k th feature :

$$y^k = \sum_{j=1}^J T_j^k w_j^k + \varepsilon^k, \quad k=1, 2, \dots, K, \quad (1)$$

where w_j^k ($j=1, 2, \dots, J$, $k=1, 2, \dots, K$) denotes the reconstruction coefficient, and ε^k denotes the reconstruction error to the k th feature of y . To some degree, reconstruction error may be caused by occlusion. Considering the possible occlusion, (1) is extended to (2).

$$y^k = \sum_{j=1}^J T_j^k w_j^k + \varepsilon_1^k + \varepsilon^k, \quad k=1, 2, \dots, K, \quad (2)$$

where ε_1^k denotes the error caused by occlusion, ε^k denotes remaining error after removing the occlusion error ε_1^k . The item ε_1^k can be represented as the combination of a group of trivial location templates $\{E_d\}$, $d=1, 2, \dots, D$, in which D is the number of trivial templates and corresponds to the size of each target template image. It

is also worth noting that each trivial template E_d is a column of identity matrix $I^{D \times D}$. Thus, (2) can be modified as (3).

$$y^k = \sum_{j=1}^J T_j^k w_j^k + \sum_{d=1}^D E_d w_d^k + \varepsilon^k, \quad k=1, 2, \dots, K, \quad (3)$$

The first two parts of (3) can be represented as a unified form, which is the linear combination of $J + D$ templates, as showed in (4).

$$y^k = \sum_{j=1}^{J+D} T_j^k w_j^k + \varepsilon^k, \quad k=1, 2, \dots, K, \quad (4)$$

where the first J templates are target templates, and the remaining D templates are occlusion templates.

Let $w^k = [w_1^k, w_2^k, \dots, w_{J+D}^k]^T$ denote the reconstruction coefficient vector of the k th feature, with entries showed in (4). Let $w_j = [w_j^1, w_j^2, \dots, w_j^k]$ denote the reconstruction coefficient vector corresponding to the j th target template or occlusion template. And let $w = [w_j^k]^{(J+D) \times K}$ denote a matrix composed of all reconstruction coefficients in (4), with the k th row and the j th column corresponding to the vector w^k and w_j respectively.

To achieve multi-cue sparse representation of all features, the proposed method constructs a cost function for each feature, as in (5).

$$\min_{w^k} \frac{1}{2} \left\| y^k - \sum_{i=1}^{J+D} T_i^k w_i^k \right\|_2^2 + \lambda \|w^k\|_1, \quad k=1, 2, \dots, K, \quad (5)$$

The cost function in (5) can be solved efficiently by the popular APG (Accelerated Proximal Gradient)method[6].

2.2 Similarity Measure

For each candidate, after the reconstruction coefficients of all features have been obtained by the previous formulas, the similarity between the candidate and target appearance model can be measured by (6).

$$RE = \sum_{k=1}^K \theta^k \|y^k - T^k w^k\|_2^2 \quad (6)$$

where, $\{\theta^k\}_{k=1}^K$ ($\sum_{k=1}^K \theta^k = 1$) are weights denoting the role of different feature in similarity measure. A simple way is to choose the reconstruction error of the feature

with minimum reconstruction error as the similarity between the candidate and the target, as in (7).

$$RE = \min_k \|y^k - T^k w^k\|_2, \quad k = 1, 2, \dots, K. \quad (7)$$

2.3 Template Update

In order to avoid and alleviate the possible drift, it is critical to adaptively update the templates using the newly obtained target. Suppose target template set at last time $t-1$ is $T_{t-1} = \{T_{t-1,1}, T_{t-1,2}, \dots, T_{t-1,J}\}$. At time t , if the tracked target \hat{x}_t with appearance y_t , is estimated with high probability (judged by a threshold), which demonstrates that the estimated target is the real target with high probability, then the tracked target is used to substitute the worst template of the template set. Otherwise the template set is not updated. Specially, after have obtained the current target \hat{y}_t at time t , (8) can be used to determine the feature i with least reconstruction error.

$$i = \arg \min_k \|\hat{y}_t^k - T_t^k w^k\|_2 \quad (8)$$

Then (9) can be used to determine the entry J of w^i with least reconstruction coefficient. Thus at last, the J th template is substituted by the target.

$$j = \operatorname{argmin}_l w_l^i \quad (9)$$

3 Tracking by Particle Filter

Particle filter[7] is a popular tracking method. In particle filter, given the target observations $\{y_1, y_2, \dots, y_t\}$ up to current time t , a set of weighted particles $\{x_{t,1}, x_{t,2}, \dots, x_{t,N}\}$ with weights $\{w_1, w_2, \dots, w_N\}$ can be used to estimate the target state \hat{x}_t of current time t in the particle filter framework. \hat{x}_t can be in the form of weighted average of all the particles as in (10).

$$\hat{x}_t = \sum_{n=1}^N x_{t,n} w_n \quad (10)$$

Or the state of the particle with maximum posterior probability as in (11).

$$\hat{x}_t = x_{t,l}, \quad \text{s.t. } l = \arg \max_i (w_i) \quad (11)$$

where w_i is the confidence weight denoting the importance of the i th particle in estimating the target. Each particle is drawn from an importance distribution, and the confidence weight w_i is proportional to its posterior probability.

$$w_i \propto P(x_{t,i} | y_{1:t}) \quad (12)$$

$P(x_{t,i} | y_{1:t})$ is determined by the motion model $P(x_t | x_{t-1})$ and the appearance model $P(y_t | x_t)$.

3.1 Motion Model

The movement of consecutive frames $P(x_t | x_{t-1})$ can be modeled by an affine image warping. Specially, given a state $x_t = \{x_t^x, x_t^y, x_t^\theta, x_t^s, x_t^\alpha, x_t^\phi\}$ at time t , where $x_t^x, x_t^y, x_t^\theta, x_t^s, x_t^\alpha, x_t^\phi$ denote x position, y position, rotation angle, scale, aspect ratio and skew respectively. Each element of x_t can be generated by random walk of corresponding value in x_{t-1} , such as in the form of (13).

$$P(x_t | x_{t-1}) = N(x_t | x_{t-1}, \Sigma) \quad (13)$$

where Σ is a diagonal covariance matrix.

3.2 Observation Model

The similarity measure showed in (7) can be used to configure observation probability of each particle as in (14).

$$P(y_i | x_i) \propto \exp(-RE_i), \quad i = 1, 2, \dots, N \quad (14)$$

where N is the number of particles.

3.3 Tracking Algorithm

Step 1. Initialization:

- 1) Manually label the initial position and scale of tracked target in 1st frame;
- 2) Generate target template set around initial target position, and extract K features for each target template. A template set is set for each feature $T^k = \{T_1^k, T_2^k, \dots, T_M^k\}, k = 1, 2, \dots, K$;

Step 2. From 2nd frame, execute the following sub-steps sequentially, till to the last frame:

- 1) Sample N particles: x_1, x_2, \dots, x_N by random walking from target state of last frame;
- 2) For each particle, extract its each feature and get corresponding appearance $y_i^k, k = 1, 2, \dots, K, i = 1, 2, \dots, N$;
- 3) For each particle y_i , get reconstruction coefficients of each feature $\{w_i^k\}, k = 1, 2, \dots, K$ as in (5) by APG method;
- 4) Get the observation probability of each particle by (14) ;
- 5) Estimate tracking target by (10) or (11);
- 6) Update target template set as in section 2.3.

4 Experiments and Analysis

To evaluate the validation of the proposed multi-cue sparse representation to visual tracking, two groups of experiments are conducted on several image sequences, taking fusing two features as example. Specially, the first experiment compared the proposed method with single feature method. And the second experiment gave the results of the proposed method in some complex environment.

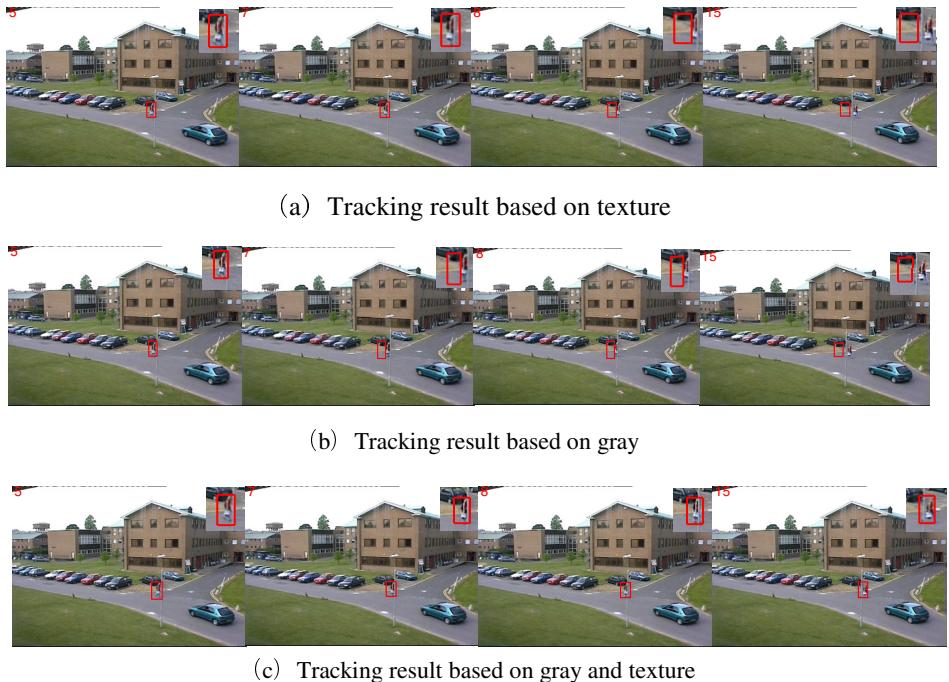


Fig. 1. The tracking result comparison between the proposed method and single-cue methods (including frames 5, 7, 8 and 15)

To validate higher performance of the proposed method compared to single feature sparse representation, we compared tracking results of the single gray feature, the single LBP texture feature and the proposed multiple features (gray + LBP texture) on PETS image sequence. In PETS sequence, a pedestrian walks on the road, during which a wire pole appears, causing occlusion. As showed in Fig1 (a) and (b), after the occlusion appeared, both single gray feature and single LBP texture feature caused drift and eventually led to failure. Looked closely, the wire pole is relatively thin and can only cause partial occlusion. During several frames of occlusion, the resolutions of gray and LBP texture varies in different frames. The resolution of gray becomes weak at some frames such as 7th frame, at this time, using only gray feature will cause drift. Likewise, the resolution of LBP texture feature becomes weak at some frames such as 8th frame, using only LBP texture feature will cause drift. However, the proposed

method fuses the two features and has following advantage: it can adaptively choose the best feature and achieve the complementarity of various features. As shown in 8th of Fig 1 (a), using LBP texture feature caused drift. And the incorporating of gray feature improved tracking accuracy, as shown in Fig 1 (c).

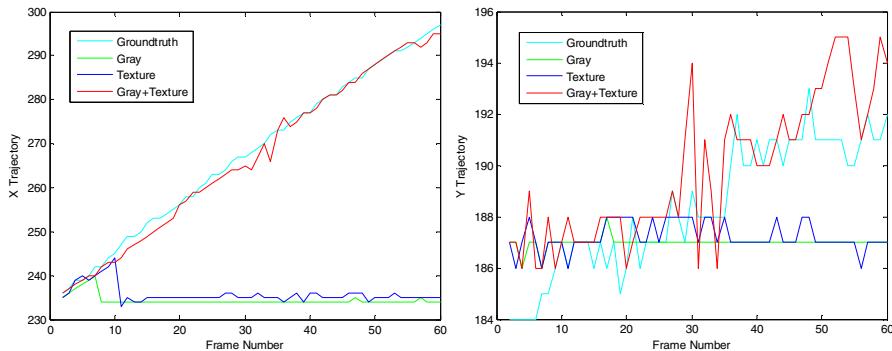


Fig. 2. The tracking trajectories comparison between the proposed method and single-cue methods



(a) Cluttered background (including frames 17,25,55 and 60)



(b) Illumination change (including frames 12,28,60 and 70)



(c) Occlusion (including frames 10,26,50 and 59)

Fig. 3. The tracking result of the proposed method in various complex environment

To quantitatively evaluate the performance of the proposed method, we compared X, Y trajectory curves of single gray feature, single LBP texture feature and the proposed method, as shown in Fig 2. From it, we can see that the proposed method performs well.

To evaluate the performance of the proposed method in various complex environments, we choose three representative image sequences: Cliffbar, Singer and Occlusion, which represents cluttered background, varying illumination and occlusion scenarios as showed in Fig 3.

5 Conclusion

A multi-cue sparse representation based visual tracking was proposed in the particle filter. The contributions are manifold: 1) Multiple features are fused and used for complementation of each other. 2) Sparse representation is embedded in the tracking framework to adaptively choose the most related templates and improve representation accuracy. 3) The extended template updating strategy is illustrated for multiple features. Experiments show that the proposed method performs more desirable with multiple complementary feature.

Acknowledgments. This research has been supported by the National Natural Science Foundation of China under the Grant Nos. 61173087 and 41071262.

References

- Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. *ACM Computing Surveys* 38 (2006)
- Mei, X., Ling, H.: Robust visual tracking using l_1 minimization. In: *ICCV* (2009)
- Kwak, S., Nam, W., Han, B., Han, J.H.: Learning Occlusion with Likelihoods for Visual Tracking. In: *CVPR* (2011)
- Liu, B., Yang, L., Huang, J., Meer, P., Gong, L., Kulikowski, C.: Robust and fast collaborative tracking with two stage sparse optimization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part IV. LNCS*, vol. 6314, pp. 624–637. Springer, Heidelberg (2010)
- Wu, Y., Ling, H., Yu, J., Li, F., Mei, X., Cheng, E.: Blurred Target Tracking by Blur-driven Tracker. In: *ICCV* (2011)
- Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Technical report (2008), <http://pages.cs.wisc.edu/~brecht/cs726docs/Tseng.APG.pdf>
- Doucet, A., de Freitas, N., Gordon, N.: *Sequential Monte Carlo Methods in Practice*. Springer, New York (2001)

LMDA: Local Maximum Discrimination Analysis

Jun Gao

School of Automation Southeast University, Nanjing, China

Abstract. In this paper, we put forward a novel supervised feature extraction method based on the Linear Discrimination Analysis: Local Maximum Discrimination Analysis. Also, in order to strengthen the local learning ability of the algorithm and improve the capable of reducing dimensionality, we introduce the Local Weighted Mean to the algorithm LMDA. Better is that, there is no Small Sample Size Problem in the new proposed algorithm with the introduction of Maximum Margin Criterion. In the end, experimental results demonstrate the above advantages of the algorithm LMDA.

Keywords: feature extraction, LWM, MMC, Small Sample Size Problem.

1 Introduction

Feature extraction as a data preprocessing method in the field of intelligent identification has been widely used. At present, Principal Component Analysis (PCA) [1] and Linear Discrimination Analysis (LDA) [2] as two classical feature extraction methods have been widely studied. PCA as an unsupervised method are mainly on the basis of the constructed covariance matrix to select the principal element. LDA as a supervised feature extraction method is to get the low-dimensional projection of the original high dimension feature space. However, dealing with high dimension small sample data, LDA will cause within-class scatter matrix to be singular, which is so-called Small Sample Size Problem (SSS). To solve this problem, researchers put forward a series of methods [3, 4], especially the methods based on Maximum Margin Criterion (MMC) [5]. These methods have relatively low time complexity [6].

LDA is using ensemble average instead of variance. However, according to statistical learning theory [7], the ensemble average partly reflects the distribution of sample information and global characteristics. LDA is partly lack of adaptability. Recently, great attention has been paid to method based on the local manifold learning. Especially, the Locality Preserving Projections (LPP) presented by He and others.

On the basis of LDA by introducing in some local learning methods [8, 9] which are successfully using Local Weighted Mean (LWM) [10], we put forward a linear discriminant analysis: Local Maximum Discrimination Analysis (LMDA). This method has the following advantages: (1) Thanks to using MMC, LMDA method avoid small sample size problem. And LMDA method bring in the QR-decomposition, which

makes the algorithm has the low time complexity;(2) We use the LWM to replace standard mean in LMDA method, which can realize preserving local neighborhood information;(3) We use manifold learning theory to divide the original sample data effectively. Thus, we improve the generalization ability of the method.

2 Related Work

2.1 Local Weighted Mean: LWM

Definition 1 (LWM) [18]. Suppose $\mathbf{X}_{1q} = \{x_{1q}^i\}_{i=1}^{n_1}$ is a local sub-domain, thus the local sub-domain \mathbf{X}_{1q} can be defined by LWM: $\sum_{i=1}^{n_1} \frac{\beta_{qi} x_{1q}^i}{\sum_{p=1}^{n_1} \beta_{qp}}$, where $0 \leq \beta_{qi} \leq 1$, and $\beta_{qi} = \exp\left(-\frac{\|x_{1q} - x_{1q}^{(i)}\|^2}{h}\right)$ is a weight parameter only related to the samples in the local sub-domain \mathbf{X}_{1q} , h^1 is the heat kernel parameter .

LWM shows the different sample contributions of keeping inner local structure through the different weights distribution of the samples in local sub-domains.

2.2 Linear Discrimination Analysis: LDA

Definition 2. Suppose $\mathbf{X} = \{x_1, \dots, x_n\}, \forall x_i \in R^d$, they are belong to C different class. Given classification decision plane normal vector ω , where within-class scatter matrix \mathbf{S}_w 、mean value u_c of class c 、between-class scatter matrix \mathbf{S}_B 、sample ensemble average u and objective function of LDA method respectively are given as follows :

$$\mathbf{S}_w = \sum_{c=1}^C \sum_{x \in D_c} (x - u_c)(x - u_c)^T, \quad \mathbf{S}_B = \sum_{c=1}^C n_c (u - u_c)(u - u_c)^T$$

$$\arg \max_{\omega^T \omega = 1} J(\omega) = \arg \max_{\omega^T \omega = 1} \frac{\omega^T \mathbf{S}_B \omega}{\omega^T \mathbf{S}_w \omega}$$

$$\text{Where } u_c = \frac{1}{n_c} \sum_{x \in X_c} x, (c = 1, 2, \dots, C), \quad U = \frac{1}{n} \sum_{x \in \mathbf{X}} x.$$

There are small sample size problems in LDA. Because of using standard mean to show variance, LDA lack the local learning ability. Therefore, we introduce LWM into the LDA and put forward LMDA method.

¹ In this paper, we make different weights of the heat kernel parameter h equal.

3 Local Maximum Discrimination Analysis: LMda

According to the manifold learning theory, the data of random distribution can be decomposed into several local sub-domains of Gaussian distribution. As shown in figure 1, we assume there are 3 class samples. For any $\mathbf{X}_c (c=1,2,3)$, we define k_c nearest neighbor data subset of x_{ci} called x_{ci} corresponding local sub-domain, which recorded as \mathbf{X}_{ci} . So we can divide the dataset \mathbf{X}_c into n_c data local sub-domains.

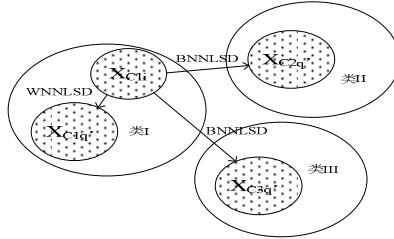


Fig. 1. LMDA basic principle diagram

We also assume there are $\mathbf{X}_{c_1}, \mathbf{X}_{c_2}$. And \mathbf{X}_{c_i} is any local sub-domain in \mathbf{X}_c . Then \mathbf{X}_{c_1q} , which meets (1) is called: the within-class nearest neighbor local sub-domains (WNNLSD) of \mathbf{X}_{c_i} in the class. Then $\mathbf{X}_{c_1q'}$, which meets (2) is called: the between-class nearest neighbor local sub-domains (BNNLSD) of \mathbf{X}_{c_i} in the class \mathbf{X}_{c_2} .

$$dis(\mathbf{X}_{c_i}, \mathbf{X}_{c_1q}) = \min_{q=1, \dots, n_c, q \neq i} dis(\mathbf{X}_{c_i}, \mathbf{X}_{c_1q}) = \min_{q=1, \dots, n_c, q \neq i} \| LWM(\mathbf{X}_{c_i}) - LWM(\mathbf{X}_{c_1q}) \|_F^2 \quad (1)$$

$$dis(\mathbf{X}_{c_i}, \mathbf{X}_{c_2q}) = \min_{q=1, \dots, n_{c_2}} dis(\mathbf{X}_{c_i}, \mathbf{X}_{c_2q}) = \min_{q=1, \dots, n_{c_2}} \| LWM(\mathbf{X}_{c_i}) - LWM(\mathbf{X}_{c_2q}) \|_F^2 \quad (2)$$

Definition 3. Assume there is a sample set $\mathbf{X} = \{x_1, \dots, x_n\}, \forall x_i \in R^d$. They are belong to C different class. $\forall \mathbf{X}_c (i=1, \dots, n_c)$ is any one of the local sub-domains of \mathbf{X}_c . Through the classification decision plane normal vector ω , the within-class scatter matrix, between-class scatter matrix and the objective function are given as following:

$$\mathbf{a}_w = \sum_{c=1}^C \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} r_{ij}^{w_c} \| LWM(\omega^T \mathbf{X}_{ci}) - LWM(\omega^T \mathbf{X}_{cj}) \|_F^2, \quad (3)$$

$$\mathbf{a}_b = \sum_{c_1=1}^C \sum_{c_2=1}^C \sum_{i=1}^{n_{c_1}} \sum_{j=1}^{n_{c_2}} r_{ij}^{b_{c_1c_2}} \| LWM(\omega^T \mathbf{X}_{c_1i}) - LWM(\omega^T \mathbf{X}_{c_2j}) \|_F^2 \quad (4)$$

$$\arg \max_{\omega^T \omega = 1} J(\omega) = (1 - \gamma) \mathbf{a}_b - \gamma \mathbf{a}_w \quad (5)$$

Where: $r_{ij}^{w_c} = \begin{cases} 1 & \mathbf{X}_{ci}, \mathbf{X}_{cj} \text{ is the WNNLSD,} \\ 0 & \text{otherwise} \end{cases}$, $r_{ij}^{b_{c_1c_2}} = \begin{cases} 1 & \mathbf{X}_{c_1i}, \mathbf{X}_{c_2j} \text{ is the BNNLSD,} \\ 0 & \text{otherwise} \end{cases}$

We simplify (3) and (4). So, we get the theorem as following:

Theorem 1. According to definition 3, (3),(4) can be simplified as following:

$$\mathbf{a}_w = \text{tr}(\omega^T \mathbf{X} \mathbf{L}_w \mathbf{X}^T \omega); \mathbf{a}_B = \text{tr}(\omega^T \mathbf{X} \mathbf{L}_B \mathbf{X}^T \omega) \quad (6)$$

Proof: Firstly, we prove $\mathbf{a}_w \cdot \text{LWM}(\omega^T \mathbf{X}_{ci}) = \sum_{m=1}^{k_c} \frac{\beta_{ci}^{(m)} \omega^T x_{ci}^{(m)}}{\sum_{p=1}^{n_c} \beta_{ci}^{(p)}}, \text{LWM}(\omega^T \mathbf{X}_{cj}) = \sum_{m=1}^{k_c} \frac{\beta_{cj}^{(m)} \omega^T x_{cj}^{(m)}}{\sum_{p=1}^{n_c} \beta_{cj}^{(p)}}$. k_c is

neighbor number; $x_{ci}^{(m)}$ and $x_{cj}^{(m)}$ are respectively the NO. m sub-domain of \mathbf{X}_{ci} and \mathbf{X}_{cj} .

If we expand the weight defined in the local sub-domains to the whole sample set \mathbf{X} , then the above two weights can be expressed as:

$$\begin{aligned} \beta_{ci} &= (\underbrace{0, \dots, 0}_{n_1}, \dots, \underbrace{\beta_{ci}^{(1)} / \sum_{p=1}^{n_c} \beta_{ci}^{(p)}, \dots, \beta_{ci}^{(n_c)} / \sum_{p=1}^{n_c} \beta_{ci}^{(p)}, \dots, 0, \dots, 0}_{n_c})^T \\ \beta_{cj} &= (\underbrace{0, \dots, 0}_{n_1}, \dots, \underbrace{\beta_{cj}^{(1)} / \sum_{p=1}^{n_c} \beta_{cj}^{(p)}, \dots, \beta_{cj}^{(n_c)} / \sum_{p=1}^{n_c} \beta_{cj}^{(p)}, \dots, 0, \dots, 0}_{n_c})^T \end{aligned}$$

Then according to $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^T \mathbf{A})$, there are :

$$\mathbf{a}_w = \text{tr}(\omega^T \mathbf{X} (\sum_{c=1}^C \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} \mathbf{R}_{ij}^{w_c} \mathbf{L}_c^{ij}) \mathbf{X}^T \omega) = \text{tr}(\omega^T \mathbf{X} \mathbf{L}_w \mathbf{X}^T \omega) \quad (7)$$

Where: $\mathbf{L}_c^{ij} = \beta_{ci} \beta_{ci}^T + \beta_{cj} \beta_{cj}^T - 2\beta_{ci} \beta_{cj}^T, \mathbf{R}_{ij}^{w_c} = \text{diag}(\underbrace{r_{ij}^{w_c}, \dots, r_{ij}^{w_c}}_n), \mathbf{L}_w = \sum_{c=1}^C \sum_{i=1}^{n_c} \sum_{j=1}^{n_c} \mathbf{R}_{ij}^{w_c} \mathbf{L}_c^{ij}.$

So \mathbf{a}_w is proofed. We can also prove \mathbf{a}_B in a similar way. $\mathbf{L}_{ij}^{ij} = \beta_{ij} \beta_{ij}^T + \beta_{kj} \beta_{kj}^T - 2\beta_{ij} \beta_{kj}^T$

$$\mathbf{R}_{ij}^{b_{ij}} = \text{diag}(\underbrace{r_{ij}^{b_{ij}}, \dots, r_{ij}^{b_{ij}}}_n), \mathbf{L}_B = \sum_{c_1=1}^C \sum_{c_2=1}^C \sum_{i=1}^{n_{c_1}} \sum_{j=1}^{n_{c_2}} \mathbf{R}_{ij}^{b_{ij}} \mathbf{L}_{c_1 c_2}^{ij}.$$

To solve the problem of complicated calculation, we take the QR-decomposition strategy ($\mathbf{X}=\mathbf{QR}$) on sample set \mathbf{X} . Then, solving (5) turn to be solving (8).

$$\underset{z \in \mathbb{Z}}{\text{argmax}} J(z) = (1-\gamma) z^T \mathbf{R} \mathbf{L}_B \mathbf{R}^T z - \gamma^T \mathbf{R} \mathbf{L}_w \mathbf{R}^T z \quad (8)$$

We can get the solution of (5) is $\omega = \mathbf{Q}z$ after solving the (8).

4 Experiment

2moons is an artificial data set of obvious nonlinear manifold structure (See Fig.2, where Fig.2 (a), (b) are the training sample and test sample), so through testing the data set, we can illustrate that the efficiency of LMDA in dealing with local manifold data. Nearest neighbor classifier is used in the process of experiment.

In the process of test, we compare the test result of LMDA with the one of LDA, set the parameter $k_c = [2, 3, 4, 5]$, $h = [2^5, 2^3, 2^1, 2^0, 2^2, 2^4, 2^6]$, and use 10 - fold cross validation.

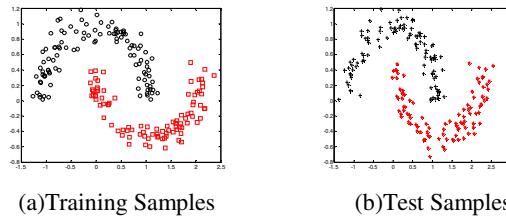


Fig. 2. 2moons Datasets

Table 1. Recognition Performance Comparison on 2moons Datasets

Datasets	LDA		LMDA	
	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
2moons	1	0.84	1	0.885(K=5;h=1;r=0.19)

According to Table1 we can see: training accuracy both can reach maximum, and test accuracy of LMDA is better than the one of LDA. This result can fully present that LMDA algorithm have a certain degree of local learning ability. It is helpful to improve the ability of feature extraction by using LWM to replace standard mean.

5 Summary

Through analysis problems in LDA, we put forward the method LMDA which has the local learning ability. We divide effectively the original sample space with the manifold learning theory, and introduce LWM into each local sub-domain. LWM can well reflect the contributions of different samples to keep the sample internal local structure and realize preserving local neighborhood information. To improve the efficiency of the algorithm, we bring in the QR-decomposition technique. Finally, the result of the test has shown that LMDA has strong local learning ability. Of course, there are still some shortages in LMDA, such as, how to more effectively improve the efficiency of this method. It will be our future research direction.

References

1. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
2. Li, R.-H., Liang, S., Chan, E.: Equivalence between LDA/QR and direct LDA. *J. International Journal of Cognitive Informatics and Natural Intelligence* 5(1), 94–112 (2011)
3. Yang, L.-P., Gu, X.-H., Ye, H.-W.: Sample locality preserving discriminant analysis for classification. *J. Optics and Precision Engineering* 19(9), 2205–2213 (2011)

4. Shu, X., Gao, Y., Lu, H.: Efficient linear discriminant analysis with locality preserving for face recognition. *J. Pattern Recognition* 45(5), 1892–1898 (2012)
5. Cui, Y., Fan, L.: Feature extraction using fuzzy maximum margin criterion. *J. Neuro. Computing* 86(1), 52–58 (2012)
6. Wan, M., Lai, Z., Jin, Z.: Feature extraction using two-dimensional local graph embedding based on maximum margin criterion. *J. Applied Mathematics and Computation* 217(23), 9659–9668 (2011)
7. Vanpanik, V.: Statistical Learning Theory. Wiley, New York (1998)
8. Lou, S., Zhang, G., Pan, H., Wang, Q.: Supervised Laplacian discriminant analysis for small sample size problem with its application to face recognition. *J. Computer Research and Development* 49(8), 1730–1737 (2012)
9. Wong, W.K., Zhao, H.T.: Supervised optimal locality preserving projection. *J. Pattern Recognition* 45(1), 186–197 (2012)
10. Atkeson, C.G., Moore, A.W., Schaal, S.: Locally weighted learning. *J. Artificial Intelligence Review* 11(1-5), 75–113 (1997)

Visual Saliency Detection via Homology Distribution and Color Contrast

Zhihui Chen, Yan Yan, and Hanzi Wang^{*}

School of Information Science and Technology, Xiamen University, Fujian, China
zhihui.qz.chen@gmail.com, {yanyan, hanzi.wang}@xmu.edu.cn

Abstract. Visual saliency detection has become a popular research topic in computer vision. It can provide useful prior knowledge for other vision tasks, such as object detection and image classification. In this paper, a new pair-wised similarity measure, *homology*, which describes how likely a pair of superpixels belongs to the same object or background, is proposed. Based on the effective combination of the homology distribution and the improved color contrast, we develop a superpixel-based saliency detection model. Compared with most of the existing models that generate fuzzy saliency maps, our model can obtain more uniformly object-highlighted maps with fewer noisy regions. In our experiments, we compare our proposed model with five state-of-the-art methods on the popular MSRA-1000 dataset. Experimental results show that the proposed model achieves a superior performance.

Keywords: Homology Distribution, Color Contrast, Visual Saliency, Superpixel-based Model.

1 Introduction

Visual saliency is a perceptual mechanism which makes an object (or a certain region of interests) stand out from background scenes so that the human perception system can rapidly focus on it. This mechanism enables human beings to get more detailed visual information that they really care about using less time. Motivated by the human perception system, visual saliency detection aims to reduce the searching space and redundant information for subsequent complex tasks. Recently, visual saliency detection has become a popular topic, due to its wide applications to the tasks of object detection, adaptive image retargeting, image segmentation, object recognition and image classification.

Previous work has explored the saliency in an image mainly based on one or several visual properties, such as contrast [1], uniqueness [2] and rarity [3]. To date, many challenging problems remain unsolved. For instance, some multi-scale models [4] generate fuzzy saliency maps, where the exact shape of a salient object cannot be well recognized. Some patch-based models [5] can not exactly match the boundary of an object or smoothly highlight the whole object.

* Corresponding author.

In this paper, we propose to detect saliency at the superpixel level, which can more accurately provide the contour of a salient object. By effectively combining two kinds of saliency properties, i.e., homology distribution and color contrast, we can obtain an accurate and uniformly highlighted saliency map. The contributions of this paper are mainly threefold. First, we make use of the concept of homology to measure the similarity between superpixels. Second, we exploit the homology distribution to describe the spatial information for both an object and the background area. Third, we combine the homology distribution with an improved variant of color contrast, which can achieve a better performance than a method using only one of them. Experimental results show that the proposed model achieves promising results and outperforms the competing methods.

The remainder of the paper is organized as follows. Related work is introduced in Section 2. In Section 3, we present the details of our proposed model including the homology distribution, improved color contrast and the combination of these two saliency properties for saliency detection. Experimental results are given in Section 4. The conclusions are given in Section 5.

2 Related Work

Based on various explanations on visual saliency mechanism, a number of saliency detection models have been proposed in recent decades. For example, Itti’s model [4] is based on center-surround contrast. It generates saliency maps for three low-level attributes at several spatial scales and then combines them to form a master saliency map. Zhang et al. [6] propose a probabilistic model called SUN (Saliency Using Natural statistics), in which saliency is defined as the self-information of visual features and learned from natural statistics. In [3], Borji and Itti use both local and global patch rarities at multiple scales as the saliency measure. Hou et al. [7] take advantage of the spectral residual feature for saliency detection.

The above models are mainly designed for predicting human fixation, which is a traditional application. However, the saliency maps are too fuzzy and insufficient to provide prior knowledge for searching and segmenting salient objects from input images. In this paper, we focus on detecting salient objects instead of predicting human fixation.

In [5], Goferman et al. develop a context-aware model which considers the dissimilarity between a patch and its K most similar patches as the saliency value. This model performs well for small-scale objects but is not suitable for large-scale salient objects.

Achanta et al. [8] present a frequency-tuned detection method, which uses the color distance between a pixel and the image mean value as the saliency measure. This method is simple but insufficient for complex natural scenes. Cheng et al. [1] propose two sparse color histogram based contrast models, including Histogram based Contrast model(HC), which computes the color contrast between each pixel and the image mean value, and Region based Contrast model (RC), which computes the color contrast between each pair of regions. Experiments show that RC performs better than HC. But RC often has some regions mixing parts of an

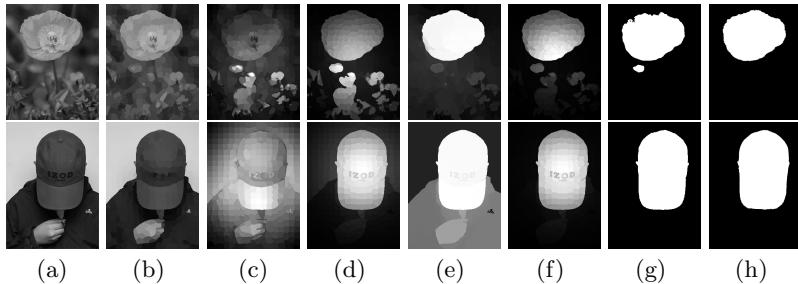


Fig. 1. An illustration of the main phases in our model: (a) Original image. (b) Superpixels of (a). (c) Traditional color contrast based saliency. (d) Improved color contrast based saliency. (e) Homology distribution based saliency. (f) Saliency maps by combining (d) and (e). (g) Adaptive-threshold segmentation of (f). (h) Ground truth mask.

object with the background, which makes the results inaccurate. Perazzi et al. [2] propose a Saliency Filters (SF) model. In the SF model, color contrast and spatial distribution are calculated respectively, and then a pixel-level smoothing is operated to obtain a full resolution saliency map. The SF model has achieved a competitive performance. However, the spatial distribution based only on color cannot handle the color variations of an appearance and may easily be distracted by background regions with similar colors.

3 The Proposed Model

In this section, we propose a superpixel-based saliency detection model by using the proposed homology distribution and the improved color contrast (Fig. 1). The proposed model involves three phases. First, we segment an input image into perceptually uniform regions which are also called superpixels. Two main advantages introduced by segmenting an image into superpixels are that the method is computationally efficient (vs. directly working on image pixels) and superpixels are inner consistent (vs. the patches used by patch based models). We use the SLIC (Simple Linear Iterative Clustering) algorithm [9] to obtain a set of approximately equally-sized superpixels in our case. Second, two complementary saliency properties, i.e., the spatial distribution and the improved color contrast, are computed respectively. The spatial distribution is based on homology, which describes the probability of two superpixels belonging to the same homogenous region. Finally, these two saliency measures are combined to generate the final saliency map.

3.1 Homology Distribution

In previous work [10,2], color distribution has been used to complement color contrast for saliency detection. Those methods decompose an input image into

a number of color components, and use the spatial distribution variance of each component as the corresponding saliency measure. However, there are at least two problems. One is that the saliency of an object would be substantially suppressed by the background region that is far away but with similar color. The other problem is the poor performance on the color-gradient area, which is usually caused by illumination and appearance changes in natural images.

To deal with the problems mentioned above, we propose a new spatial distribution measure, called *homology distribution*. Homology denotes the probability of a pair of superpixels belonging to the same object or background area. We define the pair-wise homology as:

$$H_{ij} = \exp\left(-\frac{1}{\delta_h} Dist_{ij}\right), \quad (1)$$

where $Dist_{ij}$ is the distance between two superpixels i and j ; δ_h is a scale parameter. We consider both the spatial connectivity and color similarity for the distance measure. Hence, the computation of $Dist_{ij}$ is defined as follows:

$$Dist_{ij} = \min_{path=\{sp_1, sp_2, \dots, sp_k\}} \sum_{m=sp_1:sp_{k-1}} \|F_m - F_{m+1}\|^2, \quad (2)$$

where $path$ means the least-cost path from the i -th superpixel to the j -th superpixel. $path$ consists of the adjacent superpixels from sp_1 to sp_k and $sp_1 = i$, $sp_k = j$. F_m represents the feature vector of the superpixel m . We use the CIELab color feature in the proposed model. k is the number of the superpixels along the $path$. In our experiments, $Dist_{ij}$ is computed by using a revised Floyd algorithm for efficiency.

Using the concept of homology, we calculate the homology distribution based saliency of a superpixel i as:

$$S_d(i) = \frac{1}{Z_i} \sum_{j=1:N} H_{ij} \|p_j - \mu_i\|^2, \quad (3)$$

where p_j means the position of a superpixel j ; μ_i represents the center of a homology region which the superpixel i belongs to; Z_i is a normalization factor. $\|p_j - \mu_i\|$ is the Euclidean distance between these two spatial positions. The center of a homology region μ_i can be computed as $\mu_i = \frac{1}{Z_i} \sum_{j=1:N} H_{ij} p_j$.

Compared with the traditional saliency models that simply use the spatial distribution information, our model calculates the spatial variance based on the homology which considers both color similarity and spatial connectivity in a unified way.

3.2 Color Contrast

Contrast is a widely-used saliency measure, which can be divided into two categories, that is, local contrast, such as the center-surround operation [4], and

global contrast, such as the uniform-kernel contrast [8] or the Gaussian-kernel contrast [1,2]. Local contrast often highlights the edges and texture regions. And global uniform-kernel contrast ignores the importance of context surrounding the target region. Therefore, we adopt a Gaussian kernel contrast in our model. The color contrast of superpixel i is defined as:

$$S_c(i) = \sum_{j=1:N} \omega_{ij}^p \|c_i - c_j\|^2. \quad (4)$$

where $\|c_i - c_j\|$ represents the Euclidean distance between two superpixels i and j in the CIELab color space; ω_{ij}^p is a Gaussian-kernel spatial weighted function which can be formulated as $\omega_{ij}^p = \exp(-\frac{1}{\delta_p} \|p_i - p_j\|^2)$, and δ_p is a spatial scale parameter.

Because it depends on the context information, the contrast at the center of an object is different to that on the edge. In order to uniformly highlight the conspicuous regions, a homology-based smoothing process is applied. We use homology H_{ij} as the smoothing weighted factor, which can better handle large-scale or sprawling targets. The smoothing process is formulated as:

$$\tilde{S}_c(i) = \sum_{j=1:N} H_{ij} S_c(j). \quad (5)$$

Considering that observers usually pay more attention to the center of a scene, we use a center bias mechanism in our model. Traditional center bias is built on the geometric center of an image. However, a salient object in the image would lead the observers' attention to shift. [11] describes a center-shift attention model. We apply it to filter noises far away from the target in a color contrast map. The main idea is to set the intensity centroid of the color contrast map as the shifted attention center, and build a Gaussian center bias model based on it. Therefore, the color contrast is reformulated as:

$$\hat{S}_c(i) = \tilde{S}_c(i) \exp(-\frac{1}{\delta_c} \|p_i - z_c\|^2), \quad (6)$$

where δ_c is a center bias scale parameter and z_c is the intensity centroid of $\tilde{S}_c(i)$.

3.3 Combination

As shown in Fig. 1, homology distribution and color contrast are two relatively independent and complementary saliency measures. To be specific, in an ideal case, a salient object should have a high color contrast value as well as a low homology distribution variance. To take advantage of both the measures, we combine them in a simple way as follows:

$$S(i) = \hat{S}_c(i)(1 - S_d(i)). \quad (7)$$

We normalize $\hat{S}_c(i)$ and $S_d(i)$ to the range [0,1] before the combination.

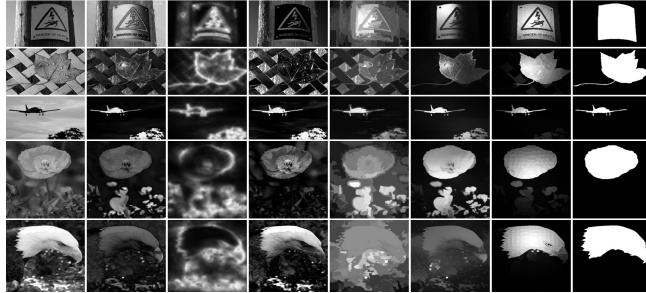


Fig. 2. Visual example results obtained by the competing methods on the MSRA-1000 database. The first column is the original image and the last is the ground-truth binary mask. The 2nd to 8th represents the saliency maps obtained by FT [8], CA [5], LC [12], RC [1], SF [2] and our proposed HCM, respectively.

4 Experimental Results

In this section, we provide comparisons between our proposed model and five state-of-the-art methods on the popular MSRA database [10]. This image database consists of 1000 images and the corresponding binary masks are provided by [8] as the ground truth. We compare the proposed Homology distribution and Color contrast based Model (HCM) with five competing models: Frequency-Tuned salient region detection (FT) [8], Context-Aware saliency detection (CA) [5], Zhai's model (LC) [12], Region-based Contrast model (RC) [1], and Saliency Filters model (SF) [2]. Fig. 2 shows visual comparisons between the competing methods.

In our experiments, δ_h is set to 200,000 for computing the homology distribution measure and smoothing the color contrast value. δ_c is set to 0.15. And δ_p is set to 0.125. The parameters are chosen empirically and fixed for all experiments.

To make comparisons, two popular evaluation criteria, Receiver Operator Characteristics (ROC) curves and Precision-Recall (PR) curves, are used to evaluate the saliency maps obtained. We also give the scores of the Area Under the Curves (AUC) as a quantitative comparison. A precision rate is the percentage of the correctly salient-assigned pixels versus all salient-assigned pixels, while a recall rate is the percentage of the correctly salient-assigned pixels versus all salient pixels in the ground truth. There is a trade-off between precision and recall. A high recall rate can be obtained at the cost of a low precision rate, and vice versa. Therefore, it is necessary to consider both of them simultaneously. Thus, the precision-recall curve is an appropriate evaluation criterion. The ROC curve is another effective and prevalent criteria for saliency model evaluation.

Evaluation with Fixed Thresholds. To compare the performance of the competing methods in details, the threshold T_0 is set to every value in [0,255] to generate binary maps from the obtained saliency maps.

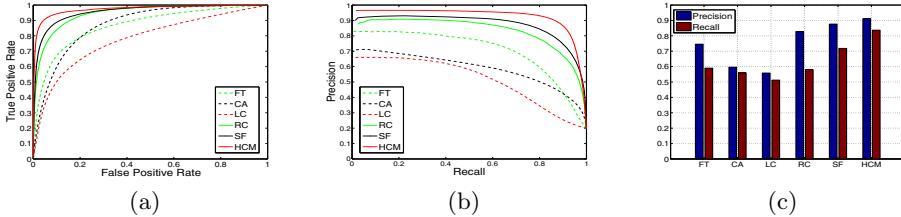


Fig. 3. Quantitative comparisons between different models. (a) The obtained Receiver Operator Characteristics curves with fixed thresholds. (b) The obtained precision recall curves with fixed thresholds. (c) The obtained precision and recall rates with an adaptive threshold. Our proposed model (*red solid*) achieves the best performance among all the competing methods.

The PR and ROC curves obtained by the six competing methods are shown in Fig. 3 (a) and (b), and the corresponding AUC scores are listed in Table 1. From Fig. 3, we can see that our proposed model achieves the best PR and ROC curves. Meanwhile, in Table 1, the proposed model achieves the highest AUC scores 0.9732 (ROC) and 0.9252 (PR). SF achieves the second highest AUC scores with 0.9561 (ROC) and 0.8718 (PR). These results prove that our model is effective and advanced due to the usage of the homology distribution and the improved color contrast.

Table 1. The AUC scores obtained by different saliency detection models

Models	$AUC_1(ROC)$	$AUC_2(PR)$	Models	$AUC_1(ROC)$	$AUC_2(PR)$
FT	0.8654	0.7052	RC	0.9451	0.8323
CA	0.8742	0.5892	SF	0.9561	0.8718
LC	0.7770	0.5151	HCM	0.9732	0.9252

Evaluation with an Adaptive Threshold. In many cases, we cannot obtain an ideal segmentation results for various images by fixing a certain threshold. Thus we also perform a quantitative evaluation for adaptive-threshold segmentation, similar to [2,8]. The adaptive threshold T_α is set as twice of the mean value of a saliency map. With T_α , we can segment a salient object from an input image adaptively. The comparison on the obtained precision and recall rates with an adaptive threshold is shown in Fig. 3 (c), from which we can see that our proposed model also achieves the highest precision and recall rates.

5 Conclusion

In this paper, we have introduced a novel pair-wise superpixel similarity measurement, *homology*, which shows a superior performance for measuring spatial saliency. We build a visual saliency detection model based on the proposed

homology distribution and improved color contrast to generate saliency maps. We compare the proposed model with five state-of-the-art models. Experimental results indicate that our model performs better than the other models.

However, the homology measure which helps to uniformly highlight a salient object and reduce the noise regions, may lead to some low-saliency object parts, especially when these parts have similar color to the background. In our future work, the shape or context information would be explored to handle this problem.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 61170179 and 61201359, by the Natural Science Foundation of Fujian Province of China under Grant 2012J05126, by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20110121110033.

References

1. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: Computer Vision and Pattern Recognition, pp. 409–416. IEEE (2011)
2. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: Computer Vision and Pattern Recognition, pp. 733–740. IEEE (2012)
3. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: Computer Vision and Pattern Recognition, pp. 478–485. IEEE (2012)
4. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), 1254–1259 (1998)
5. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: Computer Vision and Pattern Recognition, pp. 2376–2383. IEEE (2010)
6. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision* 8(7) (2008)
7. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
8. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Computer Vision and Pattern Recognition, pp. 1597–1604. IEEE (2009)
9. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. *École Polytechnique Fédéral de Lausanne (EPFL)*, Tech. Rep 149300 (2010)
10. Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
11. Yang, W., Tang, Y.Y., Fang, B., Shang, Z., Lin, Y.: Visual saliency detection with center shift. *Neurocomputing* (2012)
12. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 815–824. ACM (2006)

Multi-Modal Multiple-Instance Learning and Attribute Discovery with the Application to the Web Violent Video Detection

Shuai Hao¹, Ou Wu², Weiming Hu², and Jinfeng Yang¹

¹College of Aviation Automation, Civil Aviation University of China, Tianjin, China
shao_yjs11@yeah.net, jfyang@cauc.edu.cn

²Institute of Automation, Chinese Academy of Sciences, Beijing, China
{wuou,wmhu}@nlpr.ia.ac.cn

Abstract. Along with the ever-growing web, violent video sharing in the Internet has interfered with our daily life and affected our, especially children's health. Therefore violent video recognition is becoming important for web content filtering. In this paper, we classified the video into violent and nonviolent using Multi-Modal Multiple-Instance Learning and Attribute Discovery approach by combining audio-video with text information for web video detection. The main work is two-fold. First, we build the training bags from our video data and use attribute learning to explain attributes relation. Second, we design an efficient instance selection technique by utilizing audio-video and text information to speed up the training process without compromising the performance. The experimental results on 200 videos collected from the sharing video sites show that the proposed method is effective on violent video detection.

Keywords: Violence Detection, Multi-Modal, Multiple-Instance Learning (MIL), Attribute.

1 Introduction

The internet provides great convenience for us to obtain all sorts of information that we need. However, some violent content, including violent text, violent image and violent video etc, can also easily interfere with our daily life and affect our, especially children's health. To protect our psychological health, it is necessary to find an effective way to automatically detecting the web violent content. In this paper, we focus on the recognition of the web violent video.

Some approaches have been proposed to violent video detection. Nam et al. [1] performed violent scenes by detecting flame and blood, and various audio effects, such as gunshots, explosions. Cheng et al. [2] proposed a hierarchical approach to recognizing gunshots, explosion, and car-braking. Swanson et al. [3] implemented the automatic conversion of movies with an 'R' rating to a 'PG' rating by hiding violent scenes using video data hiding techniques. Lin et al. [4] performed violent shot by

detecting with audio and video features in movie databases. Datta et al. [5] exploited the accelerate motion vector to detect fist fighting, kicking. Giannakopoulous et al. [6] used eight audio features, both form the time and frequency domain, as input to a binary classifier which decides the video content with respect to violence. In summary, they extracted audio and video features according to some common “film grammars”, but in this paper we extract features based on attribute discovery. And they also ignored the text information, especially video introduction and user reviews that they have many information about web video content.

In this paper, we propose a novel violent video recognition model based on Multi-modal Multiple-Instance Learning and Attributes Discovery. The model assumes that, for any violent video is treated as a bag, and each bag has different number of instances (shots). Based on the existence of positive instance in the bag, we decide whether a bag is a positive training bag or not. A bag is positive bag if at least one of its instances is positive instance. And we introduce attribute discovery, discover attributes from text information and extract features based on attributes.

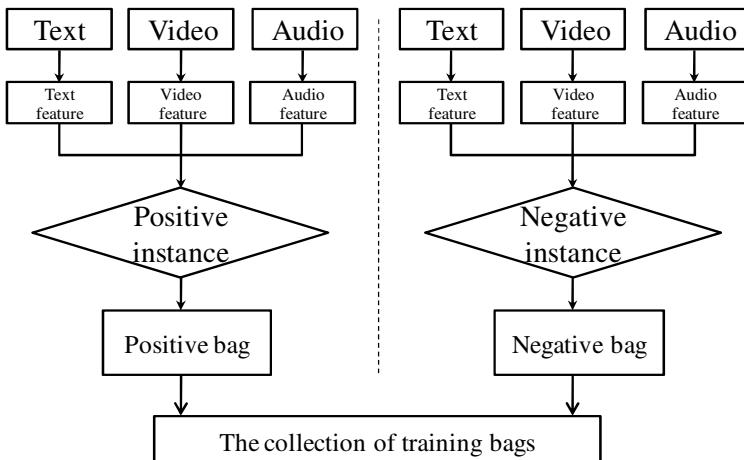


Fig. 1. The collection of training bags

The rest of this paper is organized as follows. The proposed method is described from Section 2 to 4. In the section 5, the experimental results are reported. Finally, Section 6 concludes the paper.

2 Attribute Discovery

This paper explores discovery of attribute vocabularies and learning audio-video representations from text data on the web. To begin with, we collect 180 texts associated with violent video from the Internet and rank words of text by how well their occurrence can be predicted from audio and video features. In general, possible sources of text information corresponding to web video: title, resume, user reviews,

the introduction of the video. In this paper, we put the video introduction and user reviews as the main text information that it has the user's strongly emotional expression and also have an overall description about web video. Our discovery process identifies that can be consistently predicted from some aspects of video appearance and audio representation.

2.1 Finding Video and Audio Synsets

The web data for an object category is created on and collected from a variety of internet sources (websites with different authors). Therefore, there may be several attribute phrases that describe a single same attribute. For example, “Bloody” and “Sanguinary” might be used by different sources to describe the same visual appearance characteristic of a video. Therefore, using both as attributes would be redundant. Ideally, we would like to find a comprehensive, but also compact collection of visual attributes for describing. To do so we merge attribute words based on text analysis— for example merging attributes with high co-occurrence or matching words. This process results in a collection of attribute synsets that cover the data well, but tend not to be visually repetitive. For example, Blood synsets: {“bleed”, “blood”, “bloody”, “bleeding”, “sanguinary”}, Explosion synsets: {“explosion”, “blowout” , “explode”, “exploding”, “blow-up”}.

On the text side, we keep our representation very simple. To a text, after removing stop words, and punctuation, we consider all remaining high DF (Document Frequency) words as attributes.

2.2 Attributes Relation Graph

We now describe how to build the attribute relation graph $G = \{v, \varepsilon\}$. We will assume G is a tree structured graph. A vertex $i \in v$ corresponds to the j -th attribute. An edge $(i, j) \in \varepsilon$ means the i -th and the j -th attributes have dependencies. In practice, the dependencies between certain attribute pairs might be weaker than others, i.e. the value of one attribute does not provide much information about the value of the other one. We can build a graph that only contains edges corresponding to those strong dependencies. We adopt an automatic process to build G by examining the co-occurrence statistics of attributes in the training data. First, we measure the amount of dependency between the i -th and the j -th attributes using the normalized mutual information defined as

$$\text{NormMI}(i, j) = \frac{MI(i, j)}{\min\{H(i), H(j)\}}$$

Where $MI(i, j)$ is the mutual information between the i -th and the j -th attributes, and $H(i)$ is the entropy of the i -th attribute. Both $MI(i, j)$ and $H(j)$ can be easily calculated using the empirical distributions $\tilde{p}(h_i)$, $\tilde{p}(h_j)$ and $\tilde{p}(h_i, h_j)$ from the training data.

A large $\text{NormMI}(i, j)$ means a strong interaction between the i -th and the j -th attributes. We assign a weight $\text{NormMI}(i, j)$ to the connection (i, j) , then run a

maximum spanning tree algorithm to find the edges ε to be included in the attribute relation graph G . The attribute relation graph with 11 attributes built from our training data is shown in Fig. 2.

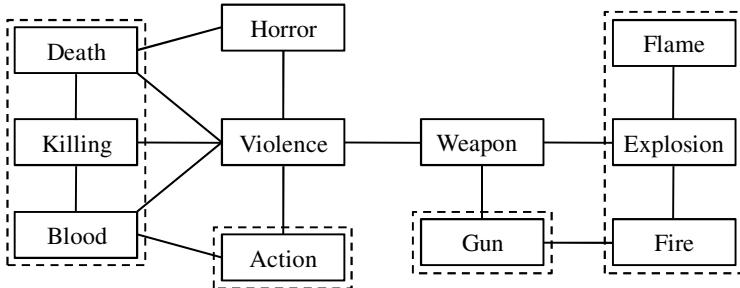


Fig. 2. Attributes relation graph

3 Features

We present in this section features generated from audio-video and text. In section 2, we have got the attributes that can be consistently predicted from some aspects of visual appearance and audio representation in violent video. Based on attribute discovery, we have got 11 attributes, but we only consider attributes: blood, action, gun that strongly predicted the visual appearance and audio representation as audio-video clues.

From the attribute blood we can infer that generally the violent video will contain bloody frame sequence, so we extract video feature bleeding. To the attribute action, we can see that the most violent video will contain high activity and abrupt motion, then we take the motion intensity as video feature. The last attribute gun that can infer the video appearance flame (fire) and the audio representation gunshot (explosion), because there will be gunshot and fire when a shot. Here, we focus on the abrupt change in energy level of the audio signal gunshot (explosion) as audio feature and flame as video feature. To effective measure this temporal signature, we use the audio energy and energy entropy criterion used in [11]. In summary, we extract video features (motion intensity[8], flame[9], bleeding[10]) and audio features (audio energy[11], energy entropy[11]) that can reflect video content well.

The text information is available for all the web videos that involves semantic information of the video content and reflects the video content, especially video introduction and user reviews. To a text, it is first preprocessed: removing noise, Chinese word segmentation and then remove stop words that frequently occur in text information and are not of contribution for classification. We adopt TF-IDF represent the text information. Finally, the value of audio-video and text feature are all normalized to be in the interval [0, 1].

4 Multiple-Instance Learning

Our classification framework is illustrated in Fig. 3 which shows the training process, where instance selection is first performed to construct instance space. Then all the training bags can be embedded into the instance space. The means that MIL problem is converted into a supervised problem via similarity based feature mapping using the selected instances. Then training bags are used to train the initial SVM classifier.

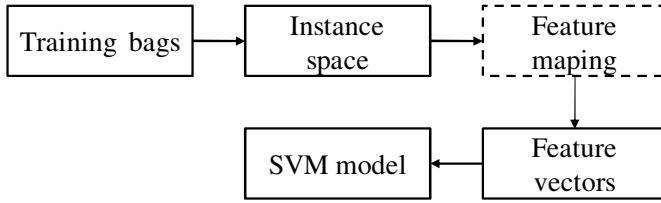


Fig. 3. The framework of our method

4.1 Notation

In the following, each webpage is treated as a bag, the training set can be denoted as $B = \{B_1^+, B_2^+, \dots, B_l^+, B_1^-, B_2^-, \dots, B_l^-\}$ where $B_i^{+/-}$ denotes the i -th bag from the positive (negative) bag and $l^{+(-)}$ denotes the number of positives (negative) bags. For the sake of convenience, we omit the sign $+/-$ when there is no need for distinction. Fig. 1 shows that collection of training bags. The label of the bag B_i is $l(B_i) \in \{+1, -1\}$. The bag $B_i = \{x_{i1}, x_{i2}, \dots, x_{in_i}\}$, the number of instances n_i varies. Each shot and text in a webpage is an instance $x_{ij} = \{v_j, a_j, w_j\}$ $j = (1, 2, \dots, n_i)$ which is consisted of the video vector $v_i = \{v_{j1}, v_{j2}, \dots, v_{jr}\}$ and the audio vector $a_i = \{a_{j1}, a_{j2}, \dots, a_{jl}\}$ and the text vector $w_j = \{w_{j1}, w_{j2}, \dots, w_{ju}\}$.The parameter r is the number of video features. The parameter l is the number of audio features. The parameter u is the number of the key words. The entire instances in the training set are represented as $x^k, k = 1, 2, \dots, n$ where

$$n = \sum_{i=1}^{l^+} n_i^+ + \sum_{i=1}^{l^-} n_i^- \quad (1)$$

4.2 Review of MILES

Our work extends ideas from MILES, so we first review it before introducing our method. MILES is mainly divided into two steps: one is instance-based feature mapping, for building instance space, the other is 1-Norm support vector machines learning, for constructing classifiers and selecting important features simultaneously. The whole instance $C = \{x^k: k = 1, 2, \dots, n\}$ in the training bag is used to build the instance space, and then all of the training bags are embedded into the instance space via a similarity between the instance x^k and the bag B_i is

$$s(x^k, B_i) = \max_j \exp\left(\frac{\|x_{ij} - x^k\|}{\sigma^2}\right) \quad (2)$$

For all the training set of l^+ positive bags and l^- negative bags, the mapping yields the following matrix representation of all training bags in the instance space

$$\begin{aligned} & [m_1^+, \dots, m_{l^+}^+, m_1^-, \dots, m_{l^-}^-] \\ &= \begin{bmatrix} s(x^1, B_1^+) & \dots & s(x^1, B_l^-) \\ s(x^2, B_1^+) & \dots & s(x^2, B_l^-) \\ \dots & \dots & \dots \\ s(x^n, B_1^+) & \dots & s(x^n, B_l^-) \end{bmatrix} \end{aligned} \quad (3)$$

After feature mapping, the MIL problem is converted to the supervised problem, then the classification problem is to find a linear classifier $y = \text{sign}(w^T m + b)$, where w and b are model parameters and m corresponds to a bag. They rewrote the parameter $\|w\|_1 = \sum_k w_k$, $|w| = u_k - v_k$, where $u_k, v_k \geq 0$. If either u_k or v_k has to equal to 0, we have $|w_k| = u_k + v_k$. The total loss function is defined as

$$\mu \sum_{i=1}^{l^+} \xi_i + (1 - \mu) \sum_{j=1}^{l^-} \eta_j \quad (4)$$

Where μ and $1 - \mu$ penalize differently on false negatives and false positives, and $0 < \mu < 1$. Then the SVM approach constructs classifiers based on hyper planes by minimizing a regularized training error.

$$\begin{aligned} & \min_{u, v, b, \xi, \eta} \lambda \sum_{k=1}^n (u_k + v_k) + \mu \sum_{i=1}^{l^+} \xi_i + (1 - \mu) \sum_{j=1}^{l^-} \eta_j \\ & [(u - v)^T m_i^+ + b] + \xi_i \geq 1, i = 1, \dots, l^+, \\ & -[(u - v)^T m_j^- + b] + \eta_j \gg 1, j = 1, \dots, l^-, \\ & u_k, v_k \geq 0, k = 1, \dots, n, \\ & \xi_i, \eta_j \geq 0, i = 1, \dots, l^+, j = 1, \dots, l^-. \end{aligned} \quad (5)$$

Where ξ, η are hinge losses, λ is called the regularization parameter. Finally, the classification of testing bag B_i is computed as

$$y = \text{sign}\left(\sum_{k \in I} W_k^* s(x^k, B_i) + b^*\right) \quad (6)$$

Let $w^* = u^* - v^*$ and b^* be the optimal solution of (5). The set of selected features is given as $\{s(x^k, \cdot) : k \in I\}$ where $I = \{k : |w_k^*| > 0\}$ is the index set of nonzero entries in W^* .

5 Experimental Result

5.1 Data Collection

Owing to a lack of open large data sets for web violent video recognition, we collected and created a video-bag set from the Internet. Each bag is composed of the video segment and text information (video introduction and user reviews). The video-bag set contains 100 violent video-bags and 100 non-violent video-bags, which have been manually labeled and segmented into shots. The normal web videos include a wide range of categories consisting of comedies, dramas, documentaries, romances and dance films. The video-bag set is from three favorite search engines: google.com, bing.com and baidu.com.

5.2 Experiments on the Dataset

The performance of the proposed approach is tested on the video dataset, and the detailed experiment result is listed in Table 1.

Three measurements, i.e. precision P , recall R and F_1 , are utilized to evaluate our system and they are defined by $P = \frac{n_{tp}}{n_p}$, $R = \frac{n_{tp}}{n_t}$, $F_1 = \frac{2*P*R}{P+R}$, where n_{tp} denotes the number of correctly detected violent videos, n_p represents the number of videos declared as violence, and n_t is the number of video labels as violence manually. F_1 – measure is a harmonic mean of precision and recall, and is used to measure the overall performance of the method.

Table 1. The result of our method in the dataset

Precision	Recall	F_1
90.11%	88.65%	89.37%

To verify the validity of the proposed approach, the comparison between SVM and our method is summarized in Table 2.

Table 2. The result of our method and SVM

Feature	SVM			Our Method(MILES)		
	Precision	Recall	F_1	Precision	Recall	F_1
T	81.71%	80.25%	80.97%	--	--	--
A+V	79.74%	80.64%	80.19%	86.55%	84.32%	85.42%
T+A+V	84.22%	83.41%	83.81%	90.11%	88.65%	89.37%

Where T represents text feature, V denotes video feature, A denotes audio feature. T means that we only extract the text feature associated web video. A+V means that we neglect the text information, extract audio-video feature. T+A+V is that using three-modal (text, video and audio) to detect violent video. Both measurements of our method are much better than those generated by SVM. We also see that the Multi-Modal is much better than single modal or bimodal generated by SVM and MILES. One possible reason should be Multi-Modal has much more information than single modal or bimodal, then we take advantage of these information through feature fusion.

Other reason is that we extract audio-video feature that have a strong relation.

6 Conclusion and Future Work

We have proposed a web violent video classification algorithm using multi-modal multiple instance learning and attribute learning approach. In our approach, text information is used design attribute discovery, we extract audio-video features based on attributes from the text information. And we collect text form video introduction and user reviews that have important information associated the video. In the future, we plan to build attributes classifier to increase the performance, and investigate method for fusion the text and audio-video features.

Acknowledgement. This work is partly supported by NSFC (Grant No. 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Nam, J., Alghoniemy, M., Tewfik, A.H.: Audio-visual content-based violent scene characterization. In: IEEE International Conference on, pp. 353–357 (1998)
2. Cheng, W.H., Chu, W.T., Wu, J.L.: Semantic context detection based on hierarchical audio models. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. pp. 109–115 (2003)
3. Swanson, M.D., Zhu, B., Tewfik, A.H.: Data hiding for video-in-video. In: IEEE International Conference on, pp. 676–679 (1997)
4. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 930–935. Springer, Heidelberg (2009)
5. Datta, A., Shah, M., Da Vitoria Lobo, N.: Person-on-person violence detection in video data. In: IEEE 16th International Conference on, pp. 433–438 (2002)
6. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D., et al. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
7. Peker, K.A., Divakaran, A., Papathomas, T.V.: Automatic measurement of intensity of motion activity of video segments. In: SPIE Conference on Storage and Retrieval for Media Databases, vol. 4315, pp. 341–351 (2001)
8. Toreyin, B.U., Dedeoglu, Y., Gudukbay, U., et al.: Computer vision based method for real-time fire and flame detection. Pattern Recognition Letters, 49–58 (2006)
9. Chen, L.H., Hsu, H.W., Wang, L.Y., et al.: Violence detection in movies. In: IEEE 2011 Eighth International Conference on, pp. 119–124 (2011)
10. Sinha, D., Tewfik, A.H.: Low bit rate transparent audio compression using adapted wavelets. IEEE Transactions on Signal Processing 41(12), 3463–3479 (1993)
11. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 663–676. Springer, Heidelberg (2010)
12. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-instance learning via embedded instance selection. J. Pattern Analysis and Machine Intelligence, 1931–1947 (2006)

Design and Implementation of a Bimodal Face Recognition System

Yong Xu^{1,2,*}, Jian Yang³, Jiajie Xu¹, Qi Zhu¹, and Zizhu Fan^{1,4}

¹ Bio-Computing Center,

Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
yongxuhitsz@163.com

² Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China

³ School of Computer Science and Technology,

Nanjing University of Science and Technology, Nanjing, China

⁴ School of Basic Science, East China Jiaotong University, Nanchang, China

Abstract. Visible light face images usually have a high resolution; however, the performance of face recognition using visible light face images is usually affected by the varying illumination. It seems that near infrared face recognition might be little influenced by the varying illumination, whereas near infrared face images usually have a low resolution and some facial marks such as scars and moles cannot be reflected by the image. In this paper, we develop a low-cost bimodal face recognition system. The system first captures the visible light and near infrared images of the face and then integrates them for face recognition. The paper also proposes a score level fusion method to combine visible light and near infrared face images for face identification. The experimental results show that the proposed method performs very well in bimodal face recognition. Moreover, the paper provides a true bimodal face image database, in which the visible light and near infrared face images are captured simultaneously.

Keywords: bimodal biometrics, feature fusion, face recognition system, TPTSR.

1 Introduction

Face recognition is a non-invasive biometrics technique, whereas the majority of the other biometrics techniques such as fingerprint recognition and palmprint recognition need the cooperation of the user [1-4]. People have been more and more interested in automatic face recognition. A number of appearance-based methods have been proposed to extract and use the holistic features of the face for face recognition [5-7]. For example, principal component analysis (PCA) and linear discriminant analysis (LDA) are two well-known and widely used appearance-based face recognition methods.

We note that in the early stage almost all the face recognition systems used two-dimensional visible light face images to perform personal authentication. Later, in

* Corresponding author.

order to simultaneously exploit the three-dimensional structure and appearance information of the face, people designed the three-dimensional (3D) face recognition system [8]. It should be pointed that the near infrared face recognition (NIFR) system also has disadvantages. For example, compared with the visible light face recognition (VLFR) system, NIFR requires that the person be closer to the system for obtaining clear near infrared face images. Moreover, near infrared face images usually have a low resolution and some facial marks such as scars and moles cannot be reflected by the image. Thus, in order to take advantages of VLFRS and NIFRS, we propose to integrate the near infrared and visible light face recognition techniques for achieving a better accuracy and more reliable performance.

The integrated near infrared and visible light face recognition is a typical bimodal problem. Processing of bimodal or multimodal information is a critical capacity of the human brain, with classic studies showing bimodal and multimodal stimulation either facilitating or interfering in perceptual processing [11]. Although there is a wealth of studies in unimodal face recognition, the recognition method for the bimodal face data have been rarely studied. We propose a fast and simple bimodal recognition method using fusion technique. Fusion stage in multimodal biometric system can be implemented in sensor level, feature level, matching score level and decision level. Since the feature set contains richer information about the raw biometric data than the match score or the decision, the fusion at the sensor or feature level is expected to provide better recognition performance. However, fusion at these two levels, respectively, is difficult to implement in practice because of the following reasons: the feature sets of multiple modalities may be incompatible, and concatenating two or more feature vectors may leads to the curse of dimensionality problem.

In this paper, we design and implement a bimodal face recognition system using feature fusion technique at matching score level. The sensors cost for our bimodal face data acquisition is very less, and they has been set up for data collection. The system first simultaneously captures the visible light and near infrared images of the face and then integrates them for face recognition. The use of the bimodal face traits allows the system to achieve a higher accuracy. Actually, our system has the merits of both the near infrared and visible light face recognition.

2 Description of the System

The function of the bimodal face recognition system is as follows. It first captures the visible light and near infrared images of the face and then uses the bimodal face traits to perform face authentication. The flowchart of the system is shown in Figures 1. The system uses a touch screen to interface with the user. The system uses two identical inside cameras to capture the face images. One camera is directly exploited to capture the visible light face image. Another camera is covered by a filter and exploited to capture the near infrared face image. Each camera has a 130 million pixel CMOS sensor. As shown earlier, the system needs to simultaneously capture the visible light and near-infrared face images. However, in the nature environment, there is no strong enough near infrared light. In order to overcome this problem, the system is equipped with 28 LED near-infrared light sources. Figure 2 shows the appearance of the bimodal face recognition system.

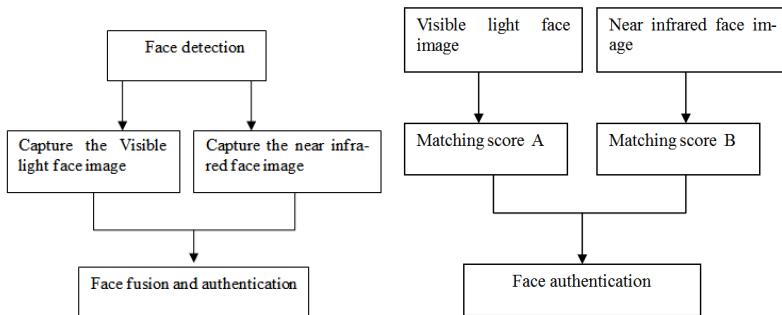


Fig. 1. The left figure shows the flowchart of the bimodal face recognition system, and the right figure shows the fusion scheme of the module “face fusion and authentication”



Fig. 2. The left figure shows the appearance of the system turned off. The right figure shows the user interface of the system.

3 The Proposed Face Recognition Method

The bimodal face recognition system first uses a modification of the two-phase test sample representation (TPTSR) method [9] to calculate the matching scores between the testing sample and training samples of the visible light face images and the near infrared face images, respectively. Then we proposed a matching score level fusion method for fusing the two kinds of matching scores to obtain final classification results. Let's have a brief overview of TPTSR. In the first step, the TPTSR seeks to construct linear representation model by all the training samples. The representation weights vector is not sparse, and it can be solved by using minimum squared error method [9]. In the first step of TPTSR, $\| y - a_i x_i \|$ is used as the measurement for evaluating the contribution of each training sample in representation. TPTSR selects M training samples that have the M greatest contributions in representing the testing sample and we denote them by x_1', x_2', \dots, x_m' . For the second step of the TPTSR, it uses the M training samples to linearly represent the testing sample by $y = a_1' x_1' + a_2' x_2' + \dots + a_m' x_m'$. The representation weights also can be calculated by MSE. For each class, get the ‘distance’ between the test sample and the combination of training samples of the i -th class determined by the first step, where the combination weights is from the combination vector obtained from the second

step. Finally, TPTSR classify the test sample y into the class that has the minimum ‘distance’ to y . We improve the TPTSR in finding the nearest neighbors of the testing sample, and introduce it in solving bimodal recognition problems. The main steps of our method are as shown below:

Step 1. Convert each face image into a one-dimensional unit vector with length of 1. For testing sample y and training sample x_i , $s_i = y^T x_i$ was used to denote the cosine similarity. The larger the s_i is, the more similar y and x_i are. The training samples that have the largest cosine similarities with the testing sample are taken as M nearest neighbors of the testing sample.

Step 2. Steps 1 and 2 of our method are implemented for visible light and near infrared face images, respectively. Let $d_1(i)$ denote the ‘distance’ between the visible light testing sample and the visible light training samples of the i -th class, determined by Step 2 of our method. Let $d_2(i)$ denote the distance between the near infrared testing sample and the near infrared training samples of the i -th class. If noone of the M nearest neighbors of the testing sample is from the j -th subject, then the ‘distance’ between the training samples of this subject and the testing sample has the maximum value among the ‘distances’ of all the subjects.

Step 3. Let $d(i) = cd_1(i) + (1-c)d_2(i)$. If $p = \arg \min_i d(i)$, then the testing sample is classified into the p -th subject.

K-fold cross validation is employed to estimate the weight c . The training data are equally divided to K subsets. In each experiment, we use K-1 folds for training set and the remaining one for validation set. The classification accuracy is the average accuracy of the K experiments. We choose the parameter c corresponding to the highest classification accuracy.

For visible light face images, if the nearest neighbor classifier and the distances between the testing sample and the training samples are directly used to perform classification, we refer to the method as VLFR using improved TPTSR. If the same method is applied to near infrared face images, we refer to it as NIFR using improved TPTSR.

4 Description of the Face Image Database

The system captured face images from 119 subjects under different lighting conditions. For example, upon the condition that both the left and the right lamps were on, each subject provided 2 to 19 visible light and near infrared face images. Figure 4 shows some visible light and near infrared face images. We used these images to conduct experiments. Every face image has the size of 200 by 200. We exploited only the 118 subjects for experiments. Each of these subjects had at least 5 visible light and near infrared face images.



Fig. 3. Some visible light and near infrared face images

5 Experimental Results

We used only the first five visible light and near infrared face images to perform experiments. For each subject, the first 4, 3 and 2 visible light and near infrared face images were used as training samples, respectively, while the remaining visible light images and near infrared images were used as testing samples. The visible light and near infrared face images were treated as the first and the second biometrics traits, respectively. We converted every face image into a one-dimensional vector and then resized it into a 4000-dimensional vector.

We also used LDA-based score fusion scheme (LSFS) and PCA-based score fusion scheme (PSFS) to perform face recognition, respectively. We implemented LSFS as follows: first, we exploited LDA to extract features from visible light and near infrared face images, respectively. Then, for each of these two kinds of face images, we computed the distance between the features of the testing sample and every training sample. Finally, for a bimodal testing sample, the simple sum rule was used to fuse its distances from each visible light and near infrared training sample. The testing sample was classified into the same subject of the bimodal training sample that has the smallest distance. Table 1 shows the classification error rate of LDA, when 4, 3 and 2 images per class are used for training, respectively. PSFS was implemented in the same way except that it used PCA rather than LDA. Table 2 shows the classification error rate of PCA, when 4, 3 and 2 images per class are used for training, respectively.

Figure 4, Figure 5 and Figure 6 show the variation of the classification error rate with the value of M, when 4, 3 and 2 images per class are used for training, respectively. In these figures, VLFR using TPTSR and NIFR using TPTSR means that the TPTSR method proposed in [9] is directly applied to visible light and near infrared face images, respectively. The improved TPTSR whose first step of TPTSR is replaced by the first step of the proposed method in section 3 is also applied to visible light and near infrared face images, respectively. We both use the proposed improved TPTSR in section 3 and TPTSR to classify the bimodal face data.

From Figure 4, our improved TPTSR has better classification performance than the TPTSR in classifying unimodal images. We also can find our method can obtain a lower classification error rate than the TPTSR, when they are applied to bimodal face images. The same comparison results can be found in Figure 5 and Figure 6. Comparing from all the Tables and Figures below, we can find our experimental results of the proposed method clearly outperforms PCA and LDA on bimodal face images.

Table 1. Classification error rate of LDA

Training samples per class	Error rate of LDA on visible light face images	Error rate of LDA on near infrared face images	Error rate of LSFS on bimodal biometrics
4	0.2797	0.3051	0.2627
3	0.3136	0.3220	0.2754
2	0.3559	0.5113	0.3023

Table 2. Classification error rate of PCA

Training samples per class	Error rate of PCA on visible light face images	Error rate of PCA on near infrared face images	Error rate of PSFS on bimodal biometrics
4	0.2712	0.3051	0.1949
3	0.3136	0.4237	0.2797
2	0.3588	0.5085	0.3588

Table 3. The performance of feature extraction method used in [10] + NN, feature extraction method used in [10] + our method and our method

Methods	Error rate on visible light face images (%)	Error rate of near infrared face images (%)	Error rate on bimodal biometrics (%)
The feature extraction method used in [10]+NN	19.49	33.90	18.64
The feature extraction method used in [10]+our method	18.64	41.53	15.25
our method	6.78	17.80	5.08

Table 4. The performance of our method and 4 different fusion methods

Methods	Error rate (%)
Our method	5.08
LDA+matching score level fusion	62.43
PCA+our method	10.17
PCA+feature level fusion	11.86
our method with future level fusion	5.93

The recognition method proposed in this paper is distinct different from the method used in [10]. Our method does not need any feature extraction for sample. Our method directly uses the pixel values of the image as the features, and [22] uses LBP code of the image as the features. For verifying whether the feature extraction method used in [10] is effective in improving the classification performance of our method, we carried the experiments shown in Table 3.

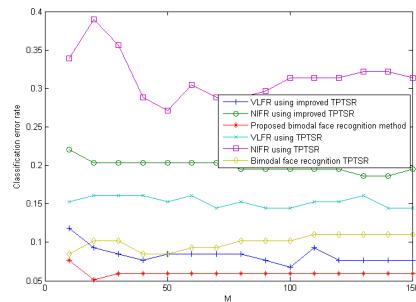


Fig. 4. The variation of the classification error rate with the value of M, when 4 images per class are used for training

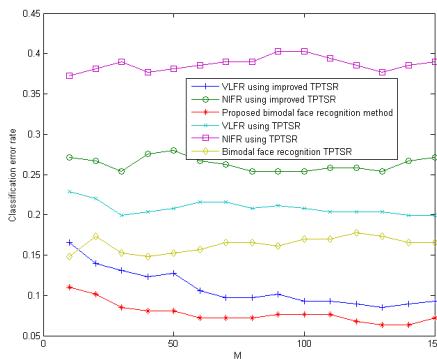


Fig. 5. The variation of the classification error rate with the value of M, when 3 images per class are used for training

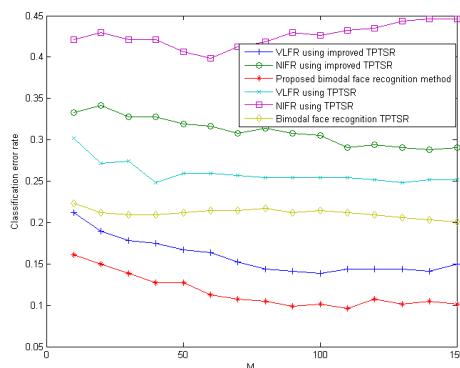


Fig. 6. The variation of the classification error rate with the value of M, when 2 images per class are used for training

We also carry some other fusion methods on the dataset, which includes LDA+matching score level fusion, PCA+our method, PCA+feature level fusion and our method with feature level fusion. The feature level fusion method used in the experiments is to directly connect the two sample vectors of the visible light and near infrared face image as one vector. Table 4 shows the results of these fusion methods

6 Conclusions

The system proposed in this paper has the merits of both the near infrared and visible light face recognition. Moreover, the simultaneously captured visible light and near infrared face images provides a true bimodal face image database, in which the visible light and near infrared face images always have the same pose and expression. Finally, the theoretical analysis show the rationale of the face recognition algorithm used. As a result, the system is able to achieve a higher accuracy.

References

1. Zhang, D., Song, F., Xu, Y., Liang, Z.: Advanced pattern recognition technologies with applications to biometrics. In: Hershey: Medical Information Science Reference (2009)
2. Jain, A.K., Flynn, P., Ross, A.A.: Handbook of Biometrics. Springer (2010)
3. Li, S.Z., Jain, A.K.: Encyclopedia of Biometrics. Springer, US (2009)
4. Jain, A.K.: Biometric authentication. Scholarpedia 3(6) (2008)
5. Xu, Y., Zhang, D.: Represent and fuse bimodal biometric images at the feature level: complex-matrix-based fusion scheme. Optical Engineering 49(3) (2010)
6. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neurosciences 3(1), 71–86 (1991)
7. Xu, Y., Zhang, D., Yang, J.-Y.: A feature extraction method for use with bimodal biometrics. Pattern Recognition 43, 1106–1115 (2010)
8. Latinus, M., et al.: Top-down and bottom-up modulation in processing bimodal face/voice stimuli. BMC Neurosci. (2010)
9. Xu, Y., Zhang, D., Yang, J., Yang, J.-Y.: A two-phase test sample sparse representation method for use with face recognition. IEEE Transactions on Circuits and Systems for Video Technology 21(9), 1255–1262 (2011)
10. Li, S.Z., Chu, R., Liao, S., Zhang, L.: Illumination invariant face recognition using near-infrared images. IEEE Trans. Pattern Anal. Mach. Intell. 29(4) (2007)

The Translation-Invariant Metric and Its Application

Bing Sun, Jufu Feng, and Guoping Wang

Key Laboratory of Machine Perception (Ministry of Education)

Department of Machine Intelligence

School of Electronics Engineering and Computer Science

Peking University, Beijing 100871, P.R. China

{bsun,wgp}@pku.edu.cn, fjf@cis.pku.edu.cn

Abstract. The image Euclidean distance (IMED) is a class of image metric that takes the spatial relationship between pixels into consideration. Sun et al. [9] showed that IMED is equivalent to a translation-invariant transform. In this paper, we extend the equivalency to the discrete frequency domain. Based on the connection, we show that GED and IMED can be implemented as low-pass filters, which reduce the space and time complexities significantly. The transform domain metric learning (TDML) proposed in [9] is also resembled as a translation-invariant counterpart of LDA. Experimental results demonstrate improvements in algorithm efficiency and performance boosts on the small sample size problems.

Keywords: IMED translation-invariant TDML.

1 Introduction

The distance measure of images plays a central role in computer vision and pattern recognition. The fact that the standard Euclidean distance assumes that pixels are spatially independent yields counter-intuitive results, e.g., a perceptually large distortion can produce smaller distance [4,11]. By incorporating the spatial correlation of pixels, two classes of image metrics, namely IMED [11] and GED [4], are demonstrated consistent performance improvements in many real world problems [4,11,1,12,14].

A key advantage of GED and IMED is that they can be embedded in any classification technique. The calculation of IMED is equivalent to performing a linear transform called the standardizing transform (ST) and then followed by the traditional Euclidean distance. Hence, feeding the ST-transformed images to a recognition algorithm automatically embeds IMED [11]. The analogous transform for GED is referred as to the generalized Euclidean transform (GET) [4].

IMED and GED are invariant to image translation. However, the associated transforms (ST and GET) are not translation-invariant (TI). This left a problem whether IMED can be implemented by a TI transform. In [9], the authors gave a positive answer to the problem and provided a proof for simple cases, yet a few technical problems are left unresolved.

In this paper, we extend the theory in [9] to the discrete frequency domain to cover the practical cases. Based on the metric-transform connection, we show that both GED and IMED are essentially low-pass filters. The resulting filters lead to the fast implementations of GED and IMED, coinciding the algorithm proposed in [8], which reduces the space and time complexities significantly. The transform domain metric learning (TDML) proposed in [9] is also resembled as a translation-invariant counterpart of LDA. Experimental results demonstrate significant improvements of algorithm efficiency and performance boosts on the small sample size problems.

The rest of this paper is organized as follows: Section 2 presents a brief review of the previous related work. In Section 3, we extend the result in [9] to discrete frequency domain. Based on the results of Section 3, we show a few applications in Section 4. Section 5 presents experimental results. Finally, a conclusion is given in Section 6.

2 A Brief Review of Previous Work

2.1 IMED and GED

The *vectorization* of an image X of size $n_1 \times n_2$ is the vector $\mathbf{x} = \text{vec}(X)$, such that the $(n_2 i_1 + i_2)$ -th component of \mathbf{x} is the intensity at the (i_1, i_2) pixel. This is a common technique to manipulate image data.

The assumption made in the standard Euclidean distance that the image pixels are spatially independent often leads to undesired results [4,11]. To solve the problem, Wang et al. [11] proposed the IMED d_G , defined as $d_G^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T G(\mathbf{x} - \mathbf{y})$. The element g_{ij} of the metric matrix G is defined by the Gaussian function [11], i.e.,

$$g_{ij} = f(\|P_i - P_j\|) = \frac{1}{2\pi\sigma^2} e^{-\frac{|P_i - P_j|^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2} e^{-\frac{(i_1 - j_1)^2 + (i_2 - j_2)^2}{2\sigma^2}}, \quad (1)$$

where $P_i = (i_1, i_2), P_j = (j_1, j_2)$.

As suggested in [11], the calculation of IMED can be simplified by decomposing G to $A^T A$. The standardizing transform (ST) is the special case when $A^T = A$, written as $A = G^{\frac{1}{2}}$. IMED can be easily embedded into most recognition algorithms. That is, feeding the ST-transformed image $G^{\frac{1}{2}}\mathbf{x}$ to a recognition algorithm automatically embeds IMED in it. Besides, Wang et al. showed that ST seems to be a smoothing [11] by illustrating a few eigen-vectors associated with the largest eigen-value of $G^{\frac{1}{2}}$.

Another image metric, called the generalized Euclidean distance (GED) [4], is essentially the same as IMED, except the distance measure between P_i and P_j . Specifically, the generating function for GED is the probability density function of the Laplace distribution $g_{ij} = e^{-\alpha \cdot (|i_1 - j_1| + |i_2 - j_2|)}$, where α is a scale parameter.

As pointed out in [11], translation-invariance (TI) is a necessary property for any intuitively reasonable image metric. Formally, for image X, Y , a distance

measure $d(\cdot, \cdot)$ is translation-invariant if and only if $d(X, Y) = d(X_\tau, Y_\tau)$, where X_τ, Y_τ is an image translation of X, Y , respectively.

Both IMED and GED depend only on the relative position between pixels P_i and P_j , i.e., $g_{ij} = g[i_1 - j_1, i_2 - j_2]$, where $i = n_2 i_1 + i_2, j = n_2 j_1 + j_2$. This makes g_{ij} invariant to image translation. However, the associated transform (ST and GET) are not TI transforms. This left a problem whether IMED and GED can be decomposed to TI transforms. That is, for any IMED or GED metric matrix G , does there exist a TI transform H such that $G = H^T H$?

2.2 The TI Transform of a TI Metric

In [9], the authors give a positive answer to the problem whether a TI metric can be implemented by a TI transform.

Theorem 1. *Given a translation-invariant metric matrix G and thus a finitely supported sequence $g[i - j] = G(i, j)$, supposing that $\hat{g}(\omega) \geq 0$ (which is the discrete time Fourier transform of $g[i]$), there exists a translation-invariant transform matrix H such that*

$$G = H^* H.$$

Theorem 1 requires $\hat{g}(\omega) \geq 0$, which is satisfied when $G \geq 0$ is an infinite-sized matrix, as a consequence of the positive operator theorem [7] or the generalized Bochner's theorem on groups [6]. In practice, G is of finite size $n \times n$. Gray [3] proved that as n approximates infinity, $\hat{g}(\omega)$ converges to a non-negative value.

The constructed translation-invariant transform matrix H is not square. Specifically, H is of size $(n + 2m) \times n$, where $[-m, m]$ is the support of the sequence $g[i]$.

3 The Discrete Frequency Domain

The authors in [9] present the theory in the continuous frequency domain, which is hardly to be applied directly in practical problems because $\hat{g}(\omega)$ is a continuous function that has to be discretized. A naive extension of Theorem 1 can be constructed by the circular convolution [5]. But there are problems.

The first problem is that the induced metric filters $g = h * h$ and $\tilde{g} = h \circledast_n h$ are different, where \circledast_n denotes the n -point circular convolution. The second problem is even worse: to derive a translation-invariant transform in discrete frequency domain, we require that the matrix representation of the metric G is circular, which is not true in both IMED and GED.

We propose the following approach to fix the arisen problems: padding the finitely supported sequences to periodic sequences. Given $h[i]$ supported on $[-m, m]$ and $x[i]$ supported on $[0, n)$, define $\tilde{h}[i]$ and $\tilde{x}[i]$ of period-($n + 2m$) by

$$\tilde{h}[i] = \begin{cases} h[i], & i \in [-m, m) \\ 0, & i \in [m, m + n) \end{cases}$$

and

$$\tilde{x}[i] = \begin{cases} x[i], & i \in [0, n) \\ 0, & i \in [-m, 0) \cup [n, m+n]. \end{cases}$$

By the circular convolution theorem [5], the two types of convolution coincide:

$$h * x[i] = \begin{cases} \tilde{h} \circledast_{n+2m} \tilde{x}[i], & i \in [-m, m+n); \\ 0, & \text{else.} \end{cases}$$

In other words, the linear convolution of h and x on its support is a period of the circular convolution of their periodic expansion \tilde{h} and \tilde{x} .

Now consider the two versions of metric filter: $g[i] = h * h^*[i]$ and $\tilde{g}[i] = \tilde{h} \circledast_{n+2m} \tilde{h}^*[i]$. Because $\forall i \in [0, n+2m), \tilde{h}[i-m] = h[i-m]$, hence $g[i] = \tilde{g}[i]$ when and only when

$$i \in [-2m, n) \cap (-n, +\infty).$$

On the other hand, by definition the metric filter is conjugate symmetric, i.e., $g[i] = \overline{g[-i]}$, $\tilde{g}[i] = \overline{\tilde{g}[-i]}$, so we assert that $g[i] = \tilde{g}[i]$ when $i \in (-n, n)$.

The above statements assert that given a finitely supported translation-invariant transform $h[x]$, the induced metric $\tilde{g}[i]$ constructed by the padded period filter $\tilde{h}[i]$ is also translation-invariant.

Hence, the analogous version of Theorem 1 is given as

Theorem 2. *Given the $[-m, m)$ supported metric filter $g[i]$, there exists a circular filter $\tilde{h}[i]$, such that $g[i]$ is equal to $\tilde{h} \circledast_{n+2m} \tilde{h}[i]$ on its support.*

Proof. Define the period- $(n+2m)$ sequence \tilde{g} by

$$\tilde{g}[i] = \begin{cases} g[i], & i \in [-m, m) \\ 0, & i \in [m, m+n]. \end{cases}$$

Let $\tilde{h}[i] = \mathcal{F}^{-1} \left(\sqrt{\widehat{\tilde{g}}[i]} \right)$ and the proof is complete. \square

It is beneficial to derive the matrix representation of Theorem 2. Given the $n \times n$ metric matrix G_n , by Theorem 1, it determines a filter $h[i]$ supported on $[-m, m)$, and hence the $(n+2m) \times n$ translate-invariant matrix $H_{m,n}$; by theorem 2, it determines a filter $\tilde{h}[i]$ of period $n+2m$, and hence the $(n+2m) \times (n+2m)$ circular matrix $\tilde{H}_{m,n}$. Writing

$$\begin{aligned} G_n &= H_{m,n}^* H_{m,n} \\ \tilde{G}_{n+2m} &= \tilde{H}_{n+2m}^* \tilde{H}_{n+2m}, \end{aligned}$$

and G_n is the left-upper $n \times n$ block of \tilde{G}_{n+2m} .

The results in discrete frequency domain can be easily extended to multi-dimensional signal space the same as in continuous frequency domain [9]. A convenient property of the extension is that the multi-dimensional data (e.g., 2d images) can be processed without vectorization.

4 Applications

4.1 The Translation-Invariant Transforms of IMED and GED

The translation-invariant transforms of IMED and GED in space and frequency domain are drawn in Fig. 4.1. It clearly shows that applying the GED or IMED is equivalent to a low-pass filtering process, which is robust to small perturbation of images.

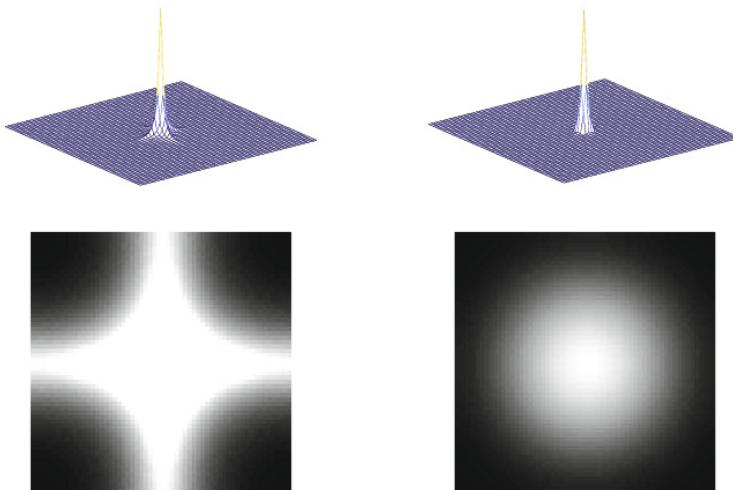


Fig. 1. The underlying filters of GED and IMED. The first row: space domain; the second row: frequency domain.

The Fast Implementation of IMED and GED. The advantages of the filtering decomposition over the GET or ST are not only the physical explanation but also the time and space complexity. Generally, the computational complexity associated with the filtering decomposition can be of $O(n \log n)$ due to the efficiency of FFT [5].

In the case of IMED and GED, since the corresponding filters decay rapidly (Fig. 4.1), we can build the period filter \tilde{g} using only several significant values. The templates of IMED ($\sigma = 1$) and GED ($\alpha = 2$) are

$$\begin{pmatrix} 0 & 0.0012 & 0.0029 & 0.0012 & 0 \\ 0.0012 & 0.0471 & 0.1198 & 0.0471 & 0.0012 \\ 0.0029 & 0.1198 & 0.3046 & 0.1198 & 0.0029 \\ 0.0012 & 0.0471 & 0.1198 & 0.0471 & 0.0012 \\ 0 & 0.0012 & 0.0029 & 0.0012 & 0 \end{pmatrix}, \begin{pmatrix} 0.0003 & 0.0025 & 0.0183 & 0.0025 & 0.0003 \\ 0.0025 & 0.0183 & 0.1353 & 0.0183 & 0.0025 \\ 0.0183 & 0.1353 & 1.0000 & 0.1353 & 0.0183 \\ 0.0025 & 0.0183 & 0.1353 & 0.0183 & 0.0025 \\ 0.0003 & 0.0025 & 0.0183 & 0.0025 & 0.0003 \end{pmatrix},$$

respectively.

Since the filter is of fixed size, the fast implementation can further reduces the space complexity from $O(n^2)$ to $O(1)$, and the time complexity from $O(n^2)$ to $O(n)$.

4.2 Transform Domain Metric Learning

Generally, in order to learn a metric G , one can do optimization with respect to G . For images of size $n_1 \times n_2$, G has $n_1^2 \times n_2^2$ elements, making the optimization intractable. Another problem is G must satisfy the positive semi-definite constraint, i.e., $G \geq 0$, so it is not easy to find efficient algorithm to solve problem with such a constraint.

With the translation-invariant assumption on G , things are much simpler. This is because the positive semi-definitive constraint $G \geq 0$ is reduced to a bound constraint $\hat{g}(\omega) \geq 0$. Furthermore, the number of parameters is the sampling number on \hat{g} , which is usually chosen to be the same as the size of input data. An additional benefit of the translation-invariant approach is that it applies to any dimensionality without modifications, thus is unnecessary to stack the multi-dimensional data to vectors.

Suppose we have some data $\{x_i\}$, and are given the data label $\{y_i\}$. Let f_i be the Fourier transform of x_i , we compute the total “similar” and “dissimilar” power spectrum:

$$p_w(\omega) = \sum_{i,j, y_i=y_j} |f_i(\omega) - f_j(\omega)|^2, \quad p_b(\omega) = \sum_{i,j, y_i \neq y_j} |f_i(\omega) - f_j(\omega)|^2.$$

The criterion here is that the filtered within-class distance is minimized, and the filtered between-class distance is maximized, simultaneously. This gives the objective functional

$$J_0(g) = \frac{\int_{T^d} \hat{g}(\omega) p_w(\omega) d\omega}{\int_{T^d} \hat{g}(\omega) p_b(\omega) d\omega}. \quad (2)$$

The objective (2) resembles the idea of LDA [2]. In fact, TDML can be viewed as a translate-invariant solution to LDA.

5 Experimental Results

In this section, we conduct several sets of experiments. The experiments are performed on 3 face data sets (UMIST, Yale and ORL database). The images in UMIST, Yale and ORL data sets are resized to 28×23 , 40×30 and 28×23 , respectively.¹ We randomly select 2 images from each class as the training set, and use the remaining images for test. We repeat the process 20 times independently and the average results are calculated.

¹ The rezization is necessary for traditional subspace and metric learning methods since they are vulnerable to the computational issue and small sample size problem from the curse of dimensionality. Our method doesn't suffer from it.

Table 1. Comparison of image metrics on various databases (%)

	ED	IMED	GED	XNZ	TDML
UMIST	60.88	60.90	62.05	60.96	73.92
Yale	71.41	71.41	71.11	67.73	75.26
ORL	81.95	81.63	80.88	81.24	84.06

We first compare TDML with several other metrics, including the standard Euclidean distance (ED), IMED, GED, and a metric learning method XNZ [13]. The performances are evaluated in terms of recognition rate using a nearest neighbor classifier. The recognition results are shown in Table. 1. TDML significantly outperforms all metrics.

Table 2. SVM classification performances of the embedded metrics. (%).

	ED	IMED	GED	TDML
UMIST	60.33	62.02	62.45	69.53
Yale	68.90	69.12	69.23	72.30
ORL	79.25	79.07	79.00	80.38

Another set of experiments was to test whether embedding the learned TI metric in an image recognition technique, e.g., SVM [10], can improve that algorithm's accuracy. Embedding a TI metric in an algorithm is simple: first, transform all images by the corresponding TI transform, and then run the algorithm with the transformed images as input data.

Table. 2 gives the results of the metric when embedded to SVM. It can be found that TDML improves the performance of SVM better than IMED and GED.

6 Conclusion

In this paper, we extend the equivalency in [9] to the discrete frequency domain. We show that GED and IMED are low-pass filters, resulting in fast implementations which reduce the space and time complexities significantly. The transform domain metric learning (TDML) proposed in [9] is also resembled as a translation-invariant counterpart of LDA. Experimental results demonstrate significant improvement of algorithm efficiency and performance boosts on small sample size problems.

One possible future direction is the search for more effective metric learning algorithm. TDML is a simple and intuitive attempt and we expect novel methods that combine the concepts of margins, kernels, locality and non-linearity.

Acknowledgments. This work was supported by NBRPC (2011CB302400).

References

1. Chen, J., Wang, R., Shan, S., Chen, X., Gao, W.: Isomap based on the image euclidean distance. In: 18th International Conference on Pattern Recognition ICPR 2006, vol. 2, pp. 1110–1113 (2006)
2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience (2000)
3. Gray, R.M.: Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory* 2(3), 155–239 (2006)
4. Jean, J.S.N.: A new distance measure for binary images. In: 1990 International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1990, pp. 2061–2064 (April 1990)
5. Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing, 2nd edn. Prentice Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs (1999)
6. Rudin, W.: Fourier Analysis on Groups. Wiley (January 1990)
7. Rudin, W.: Functional Analysis, vol. 2. McGraw-Hill Book Company, New York (1991)
8. Sun, B., Feng, J.: A fast algorithm for image euclidean distance. Chinese Conference on Pattern Recognition, CCPR 2008, 1–5 (2008)
9. Sun, B., Feng, J., Wang, L.: Learning IMED via shift-invariant transformation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1398–1405 (2009)
10. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
11. Wang, L., Zhang, Y., Feng, J.: On the euclidean distance of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8), 1334–1339 (2005)
12. Wang, R., Chen, J., Shan, S., Chen, X., Gao, W.: Enhancing training set for face detection. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 477–480. IEEE Computer Society, Washington, DC (2006)
13. Xiang, S., Nie, F., Zhang, C.: Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition* 41(12), 3600–3612 (2008), <http://www.sciencedirect.com/science/article/B6V14-4SM62BD-2/2/2bfd57fda7833c560424f29a5b97c97c>
14. Zhu, S., Song, Z., Feng, J.: Face recognition using local binary patterns with image euclidean distance. In: SPIE, vol. 6790 (November 2007)

A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks Based Large Vocabulary Chinese Speech Recognition

Xiangang Li, Yuning Yang, and Xihong Wu*

Speech and Hearing Research Center,
Key Laboratory of Machine Perception (Ministry of Education), Peking University,
Beijing, 100871, China
`{lixg,yangyn,wxh}@cis.pku.edu.cn`

Abstract. This paper compared the performance of different acoustic modeling units in deep neural networks (DNNs) based large vocabulary continuous speech recognition (LVCSR) systems for Chinese. Recently, the deep neural networks based acoustic modeling method has achieved very competitive performance for many speech recognition tasks, and has become the focus of current LVCSR research. Some previous work have studied the context independent and context dependent DNNs based acoustic models. For Chinese, a syllabic language, the choice of basic modeling units under the background of DNNs based LVCSR systems is a very important issue. In this work, three basic modeling units: syllables, Initial/Finals, phones, are discussed and compared. Experimental results showed that, in the DNNs based systems, the context dependent phones obtain the best performance, and the context independent syllables have the similar performance with the context dependent Initial/Finals. Besides, how the number of clustered states impact on the performance of DNNs based systems is also discussed, which showed some different properties from the GMMs based systems.

Keywords: Deep neural networks, Chinese automatic speech recognition, acoustic modeling units.

1 Introduction

Although the last decades have witnessed significant progress in automatic speech recognition (ASR), the performance of ASR systems in many real usage scenarios still lags far behind human level performance. Many new machine learning algorithms have led to significant advances in ASR. Recently, a major advances have been made in training deep neural networks (DNNs), which contain more than one layer of hidden units between the inputs and outputs [4]. DNNs have been applied successfully in a number of tasks, such as the classification for images, text, and movies. The great success of DNNs has triggered the interest in developing DNNs based acoustic models in ASR.

* Corresponding author, IEEE Senior member.

Acoustic modeling is a fundamental problem in ASR. Almost all of the state-of-the-art ASR systems are hidden Markov model (HMM) based. The relationship between HMM states and the acoustic input is usually represented by Gaussian Mixture Models (GMMs) or Artificial Neural Networks (ANNs). However, the ANNs were typically trained with only one hidden layer. It has long been suspected that deep networks could model complex higher statistical structure effectively until recently many new algorithms were developed for training deep models. One of these approaches is the deep belief network (DBN) training algorithms, suggested in [2], in which, the weights of each layer were first initialized by a purely unsupervised way and then fine-tuned with the labeled data. Many researches indicated that DNNs based acoustic models can outperform GMMs in many speech recognition tasks [1]. As first introduced in [3,4], the context independent (CI) pre-trained DNN/HMM hybrid architectures have been proposed for phone recognition. Then, context dependent (CD) pre-trained DNN/HMM for large vocabulary speech recognition is studied and discussed in [5,6]. DNNs based ASR systems achieved very competitive performance, and have become the focus of current ASR research.

Deep belief network pre-training was the first pre-training method to be widely studied, and further research indicated that they could be trained in many different ways, such as the discriminative pre-training method introduced in [10], and generative pre-training with various types of auto-encoder. For fine-tuning, many alternative methods can be applied, such as stochastic gradient descent, nonlinear conjugate-gradient, LBFGS, and “Hessian-free” method. Moreover, there are many neural network architectures designed for DNN/HMMs in ASR, such as deep tensor networks [7], deep stacking networks [9], deep convex networks [8]. There are several issues about the use of DNN/HMMs in ASR need further explorations, including the choice of modeling units for some specific language, the assessment of the models on large real-world datasets, and the adaptation criteria for this kind of models.

This paper focus on the choice of acoustic modeling units in DNNs based large vocabulary continuous speech recognition systems for Chinese. In the GMMs based Chinese ASR systems, there are many researches in literature discussed the modeling units [11]. Most Chinese ASR use initial/finals (IFs) as basic acoustic modeling units set, which is mainly due to the low complexity of modeling and fair requirement on the amount of training data. Besides, some researches decompose the finals, in which, phones are adopted as the basic modeling units. Moreover, Chinese is naturally a syllabic language, thus some efforts are made to build the syllable based acoustic models. There are many studies and discussions on the Chinese acoustic modeling units on the background of GMM/HMMs ASR systems. In this work, we will report the study on how the performance of DNNs based ASR systems is affected by the amount of different acoustic modeling units: CI IFs, CD IFs, CI phones, CD phones, CI syllable and CD syllable. Besides, how the number of clustered states impact on the performance of DNNs based systems is also discussed.

The remainder of this paper is organized as follows. The next section presents the basic framework of DNN/HMMs acoustic modeling. Section 3 describes the acoustic modeling units for Chinese speech recognition in detail, followed by the experiments and results in section 4. The discussions and conclusions are drawn in the last section.

2 The Deep Neural Network HMMs

A DNN is a feed-forward, artificial neural network with many hidden layers.

DNNs can be discriminatively trained by back propagating (BP) derivatives of a cost function. For large training sets, the stochastic gradient descent method is always employed with a “momentum” coefficient. However, DNNs with many hidden layers are always hard to optimize. BP can easily get trapped in poor local optima from a random starting point. This optimization challenge can be somewhat alleviated by introducing the pre-training procedure.

The most important advance in learning for deep networks has been the development of layer-wise unsupervised pre-training methods, first provided by [12] was based on restricted Boltzmann machine (RBM). Once an RBM is trained on data, then the hidden activation probabilities of current RBM are used as the training data for the next RBM. Thus each RBM weights can be used to extract features from the output of the previous layer. When the pre-training complete, a randomly initialized softmax output layer will be added. Finally, the whole network can be fine-tune by BP. This RBM based pre-training method can be viewed as a generative pretraining method, in which, the structure of input data is firstly learned before the multi-classification task.

There are some discriminative pre-training methods. The DNN is trained by starting with a one-hidden-layer neural network. Once the networks has been trained discriminatively, a second hidden layer is interposed between the last hidden layer and the softmax output layer, and the whole networks is again discriminatively trained. This can be continued until the desired deep structure is reached, and then fine-tune to convergence by BP.

In the DNN based acoustic models, the DNN outputs the posterior probabilities of the acoustic modeling units over the input acoustic feature. In the HMM framework, the acoustic model is always formulated as:

$$p(x|w) = \max_q \pi(q_0) \prod_{t=1}^T a_{q_{t-1}} a_{q_t} \prod_{t=0}^T p(x_t|q_t) \quad (1)$$

In the GMMs based ASR systems, observation probability $p(x_t|q_t)$ is directly modeled by GMMs. However, in the DNNs based ASR systems, the observation probability $p(x_t|q_t)$ is converted by $p(x_t|q_t) = p(q_t|x_t)p(x_t)/p(q_t)$, and $p(q_t|x_t)$ is the posterior probability modeled by DNNs.

The DNNs based acoustic models can be trained using the embedded Viterbi algorithm with GMMs seeding. The GMMs based system is firstly built, then conduct a forced alignment procedure with the GMM/HMMs. The modeling

units of the GMMs are delivered as the modeling units for DNNs. Through the forced alignment, the input acoustic features are labeled, and then, the pre-trained neural net is fine-tuned discriminatively with BP.

3 Acoustic Modeling Units Selection for Chinese Speech Recognition

Chinese is naturally a syllabic language and each basic language unit can be phonetically represented by a syllable. Among the 1254 distinct syllables, there are 408 toneless base-syllables. However, in this paper, we only focus on recognizing the 408 base-syllables, thus the tone information is not discussed. In order to conduct speech recognition, the syllable is always decomposed into initial and finals, and furthermore, the finals can be decomposed into medial, main vowel and nasal three parts if necessary.

It is important to select appropriate basic units to represent acoustic information for a specific language in designing ASR systems. For Chinese, the syllable contains the most strong co-articulation, and syllable based model has little problem on constructing the lexicon for new task. However, the syllable based models suffer from the poor coverage and distribution unevenness of training data. Besides, in the IFs based ASR systems, the complexity of model is low and the requirement of data for each unit can be easily satisfied. Thus, this kind of modeling units is widely used in the Asian community. However, some finals may have much more complex acoustic representation than others, such as “iong”, which have the medial part “i”, vowel part “o” and nasal part “ng”. Therefore, many researchers have discussed the phone based systems, in which, “iong” is modeling with three phones: “i”, “o”, “ng”. The acoustic modeling units in phone based systems are very similar with the International Phonetic Alphabet units, which makes these kind of systems can be easily used for the multi-language ASR systems.

Under the background of DNNs based ASR systems, the CI and CD modeling units have been discussed and compared in literature. In the CD DNN systems, the clustered states are used as the label for the neural networks. For Chinese, the syllable, IFs and phone based models have not been systematic compared. Thus in this paper, we conducted experiments for different modeling units in Chinese ASR, which will be presented in the following section.

4 Experiments and Results

4.1 Experiments Setup

We carried out speech recognition experiments on Hub4 Chinese broadcast news database. The training set is 1997 Chinese broadcast news speech corpus (Hub-4NE) training data which contains about 30 hours of speech. The test set is Chinese broadcast news evaluation data which consist of about one hour speech.

The acoustic model training set was also used to train a 3-gram language model used for these experiments.

For the feature extraction in the experiments, the speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. In the GMMs based experiments, the speech was represented using 12th-order Mel frequency cepstral coefficients and energy, along with their first and second temporal derivatives. Channel normalization is applied using cepstral mean normalization over each utterance. In the DNNs based experiments, the speech was based on a Fourier-transform-based filter-bank with 21 coefficients distributed on a mel-scale (and energy) together with corresponding first and second order temporal derivatives. In the experiment, a context of 7 frames were used with current frame, forming a total of 945 (15×6) inputs to the DNNs.

The GMMs based acoustic models were trained using ML criteria, and contain 32 Gaussians. The DNNs used have 4 hidden layers with 2500 nodes in each layer. The DNNs were trained from the alignments with the GMMs based models. For fine-tuning, we used stochastic gradient descent with mini-batch of 128, the learning rate started at 0.006. At the end of each epoch, if the substitution error on the development set decreased less than 0.1, the learning rate begin to halving. This continued until the substitution error on the development set increased.

4.2 Comparison of Different Kinds of CI Acoustic Modeling Units

Firstly, the experiments are conducted to compare the performance of CI phones, CI IFs and CI syllables. The CI phones and CI IFs are 3 state left-to-right HMM, and the states number of each HMM for CI syllables is determined by the corresponding number of phones, for example, “qiong” have “q”, “i”, “o”, “ng” 4 phones, the states number is 7($3 + 4$); “a” have only 1 phone, the states number if 4($3 + 1$). The experimental results are placed in table 1.

Table 1. Character error rate of different context independent models

	CI-Phones(%)	CI-IFs(%)	CI-Syllables(%)
GMMs	34.27	31.23	29.95
DNNs	22.40	22.71	20.03

From table 1, we can find out that, the CI syllables got the best performance. However, the CI syllables contains more than 2000 states in total, the description of acoustic representation is more detailed, which may explain the differences of performance among these three acoustic modeling units.

4.3 Comparison of Different Kinds of CD Acoustic Modeling Units

Secondly, some experiments are conducted for the comparison of different kinds of CD acoustic modeling units. Just like the method mentioned in [CDDNN], the

context-dependent acoustic models are based on decision tree based tying. More specifically, the CD phones, CD IFs, CD syllables are modeled by tri-phones with around 4000 shared states (senones). These senones are the labels for DNNs and the modeling units for GMMs. The experimental results are placed in table 2.

Table 2. Character error rate of different context dependent models

	CD-Phones(%)	CD-IFs(%)	CD-Syllables(%)
GMMs	24.97	26.27	30.39
DNNs	18.46	20.35	19.81

Table 2 shows that the CD phones outperformed than the other two types of modeling units. However, the CI phones are much worse than other CI models, but while taking into account the context dependency, the phones become the best modeling units. The introducing of context dependency makes the phones easy to discriminate, while the context independent phones make it easier to share phonetic units across different syllables but lack of the influence caused by different syllable and the information about co-articulation.

Compared with CD IFs and CD syllables, there is remarkable gap between the GMMs system for these two types of modeling units, but performances of DNNs for these two are quite close. However, GMMs models the distributions of each senone, while DNNs care about how to classify around these senones, which makes the GMMs based systems much more easily been influenced by data coverage. In the CD syllables based systems, some syllables may have only less than 10 examples, which resulting in serious data coverage problems.

Besides, someone may point out that, unlike the other two kinds of modeling units, the syllables have not yet been improved from CI to CD. A reasonable explanation to the experimental facts may be the data coverage of these kinds of models. If there are enough data for each syllable, the performance may improve further.

4.4 The Impact of the Number of Senones

Nevertheless, the performance is quite affected by the data size, thus, while discussion about the acoustic modeling units, the number of senones should been taken into account. Thus, we have conducted some experiments on the CD phones, in which, GMMs and DNNs based on different number of senones were trained and tested. The results are showed in Fig. 1.

From Fig. 1(a), we can find out that, the ASR performance varied with the number of senones. However, for GMMs based acoustic models, the performance becomes worse when the number increase, while for the DNNs based acoustic models, the CER maintains about 18.50%. The frame classification accuracy of the training and developing set is represented in Fig. 1(b), which shows that the classification becomes harder when the number of senones increases. Compared

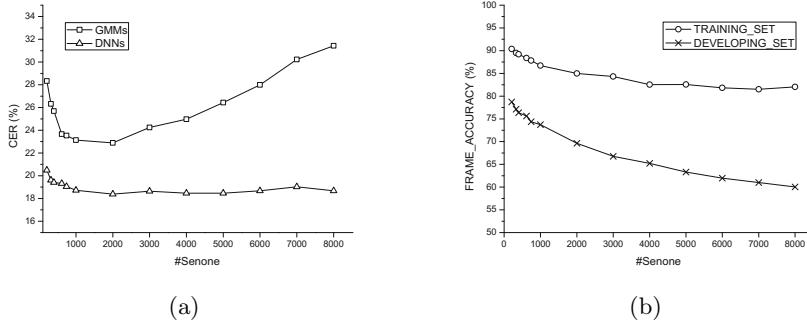


Fig. 1. The performance varying with the number of senones. (a) compare The character error rate of DNNs and GMMs based systems. (b) compare the frame classification accuracy on the training and developing set varying with the number of senones.

with these two figure, the conclusion is that when the number of senones increase, the classification performance of DNNs and the ASR performance of GMMs will decrease, but the ASR performance of DNNs based acoustic models has been less affected. ASR is a sequences pattern recognition problem, where many knowledge resources are intergraded, such as language models and dictionaries. Due to the constraint of language model, the classification performance decrease of DNNs has not result in the significantly decrease of the performance of ASR.

5 Discussions and Conclusions

This paper presents a systematic performance comparison among various levels of acoustic modeling units for DNNs and GMMs based Chinese speech recognition. For the context independent acoustic modeling units, syllable based models have showed better performance than Initial/Finals or phone based models, especially in the DNNs based ASR systems. The outstanding performance mainly benefits from the more detailed description of acoustic representation. In addition, the best performance is obtained with the context dependent phones in the DNNs systems. When the context dependency information is introduced, the performances of Initial/Finals and phones have gained remarkable improvement. Besides, for the DNNs based systems, the impact of the number of senones is also discussed. Unlike the GMMs based systems, when the number of senones increases, although the classification performance of DNNs decrease, the ASR performance of DNNs based acoustic models has been less affected. What should be pointed out is that, with DNNs, the context independent syllable based systems have gain the similar performance with context dependent Initial/Finals based systems.

DNNs based acoustic models showed many important properties. The performance would not decrease seriously facing the distribution unevenness of training data. There is little impact of senones number on the performance of DNNs based ASR systems.

While introducing the DNNs into the Chinese speech recognition, the performance have obtained a great improvement. Compared with the best performance of GMMs based systems, the DNNs can obtain more than 20% relative character error rate decrease. We believe this work on DNNs based Chinese speech recognition is only the first step towards a power Chinese speech recognition systems. There are many efforts need to be done, more specifically, the clustering strategy, the tone modeling, the neural networks structures and so on.

Acknowledgments. The work was supported in part by the National Natural Science Foundation of China (No. 91120001, No.61175043), and a “Twelfth Five-Year” National Science & Technology Support Program of China (No. 2012BAI12B01).

References

1. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Mag.* 29(6), 82–97 (2012)
2. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. *Neural Computation* 18(7), 1527–1554 (2006)
3. Mohamed, A., Dahl, G., Hinton, G.: Deep belief networks for phone recognition. In: Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications (2009)
4. Mohamed, A., Dahl, G., Hinton, G.: Acoustic modeling using deep belief networks. *IEEE Trans. Audio Speech Lang. Processing* 20(1), 14–22 (2012)
5. Dahl, G., Yu, D., Deng, L., Acero, A.: Context-dependent pretrained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Processing* 20(1), 30–42 (2012)
6. Seide, F., Li, G., Yu, D.: Conversational speech transcription using context-dependent deep neural networks. In: Proc. Interspeech, pp. 437–440 (2011)
7. Yu, D., Deng, L., Seide, F.: Large vocabulary speech recognition using deep tensor neural networks. In: Proc. Interspeech (2012)
8. Deng, L., Yu, D.: Deep convex network: A scalable architecture for speech pattern classification. In: Proc. Interspeech, pp. 2285–2288 (2011)
9. Deng, L., Yu, D., Platt, J.: Scalable stacking and learning for building deep architectures. In: Proc. ICASSP, pp. 2133–2136 (2012)
10. Yu, D., Deng, L., Li, G., Seide, F.: Discriminative pretraining of deep neural networks. U.S. Patent Filing (November 2011)
11. Wu, H., Wu, X.H.: Context Dependent Syllable Acoustic model for Continuous Chinese speech recognition. In: Proc. Interspeech, pp. 1713–1716 (2007)
12. Hinton, G.: A practical guide to training restricted Boltzmann machines. Technical Report UTML TR 2010-003, University of Toronto (2010)

Multi-level Linguistic Knowledge Based Chinese Grapheme-to-Phoneme Conversion

Yi Liu, Xiaojun Chen, Caixia Gong, and Xihong Wu*

Speech and Hearing Research Center, Key Laboratory of Machine Perception
(Ministry of Education), School of Electronics Engineering and Computer Science,
Peking University, Beijing, P.R. China
`{liuy, chenxj, gongcx, wxh}@cis.pku.edu.cn`

Abstract. This paper proposes a novel method integrating multi-level linguistic knowledge for Chinese grapheme-to-phoneme(G2P) conversion. Pronunciation prediction of non-standard words(NSWs) and disambiguation of polyphonic characters are two important issues in Chinese grapheme-to-phoneme conversion. Considering effect of linguistic knowledge, multi-level linguistic cues, including word form, Part-of-Speech (POS), named entity, collocation and syntactic structure, are extracted under a unified syntactic parsing framework and integrated by maximum entropy approach to disambiguate polyphonic characters. Besides, the text normalization is incorporated in this framework to help predict pronunciation of non-standard words. Experiment results show that the proposed method can improve the performance from 95.64% to 99.23%.

Keywords: Chinese grapheme-to-phoneme conversion, multi-level linguistic knowledge, non-standard words, polyphonic characters, syntactic structure.

1 Introduction

An accurate G2P conversion is required for various speech processing tasks, such as automatic speech recognition (ASR) and text-to-speech (TTS). For ASR, speech transcriptions which are always word or grapheme sequences need to be converted to phoneme sequences in the training stage, and it is important to determine the pronunciation of lexicon in the decoding stage. And for TTS, predicting the proper pronunciations of a grapheme sequence directly affects the intelligibility of synthetic speech. Therefore, G2P conversion is a critical component in both ASR and TTS.

G2P conversion is the task of finding pronunciations of a word sequence given its orthographic form. It can be formalized using Bayes' decision rule as

$$\varphi(g) = \operatorname{argmax}_{\varphi' \in \phi^*} p(\varphi', g). \quad (1)$$

* Corresponding author, IEEE Senior member.

This means, for a given orthographic form (grapheme sequence), $g \in G^*$, seeking the most likely pronunciations (phoneme sequence) $\varphi \in \phi^*$.

In Chinese G2P conversion, two important issues including both uncertainty of non-standard words (NSWs) and Chinese polyphonic ambiguity affect its performance. Firstly, NSWs contain different pronunciations according to different meanings. For instance, in sentence “中国民间艺人制作了2008个宫灯迎接2008年北京奥运会”，the first “2008” means quantity while the second “2008” represents the year of 2008 indicating time. Secondly, among 6763 frequently used Chinese characters, 928 characters have more than one pronunciation, which are also called polyphonic characters [1]. Due to the impact of multi-level linguistic cues, it is difficult to determine their pronunciations through simply looking up a lexicon. Hence rich linguistic cues, such as word form, POS, named entity and syntactic structure are required.

So far, approaches proposed to solve Chinese G2P conversion can be divided into three categories: rule-based, statistics-based and a combination of both. In rule-based methods, many cues including POS, surname[2], prosodic words[3] and so on, are exploited to get these rules handcrafted by language experts. But it is time-consuming to summarize and maintain the large set of rules, and conflicts between them appear frequently along with their increasing. With the availability of annotated corpora, many statistics-based methods, such as decision tree[4], N-gram[5], stochastic decision list[6] and maximum entropy[7], are developed, which treat Chinese G2P conversion as a classification task. In these methods, a statistical model is firstly trained by extracting lexical features and then applied to predict the pronunciations of polyphonic characters. However the accuracy of these methods depends on both large-scale annotated corpus and use of multi-level linguistic cues. To take advantage of these two methods, the combination of them is proposed to enhance the performance[8,9,10]. However some drawbacks exist in these methods. Firstly, the linguistic cues are extracted based on the Chinese lexical analysis which are often realized in a cascade manner. Its disadvantage is that errors generated by earlier subtasks propagate through the pipeline and will never be recovered in downstream subtasks. Moreover, this manner prevents information sharing among this cascade procedure. Secondly, the pronunciations of NSWs, which also affect Chinese lexical results, are seldom dealt with in traditional methods. Thirdly, some important syntactic cues are still not considered in existing methods.

In this study, pronunciation prediction of NSWs and disambiguation of Chinese polyphonic characters are integrated into a unified framework. In addition to the lexical-level cues, such as word form, POS and named entity, the syntactic-level cues, including syntactic structure and collocation, are introduced to disambiguate polyphonic characters. To prevent error propagation in the cascade procedure, word segmentation, POS tagging and named entity recognition are integrated into a unified syntactic parsing framework. Moreover, pronunciation prediction of NSWs is also incorporated into this framework. Finally, these multi-level linguistic cues are integrated using maximum entropy approach to disambiguate polyphonic characters.

This paper is organized as following: Section 2 investigates the two crucial issues of Chinese G2P conversion. The proposed method is introduced in Section 3. Furthermore, experimental results are shown and analyzed in Section 4. The conclusions and future works are listed in Section 5.

2 Two Crucial Issues for Chinese G2P

2.1 Non-Standard Words

In Chinese sentences, there are many NSWs, including numbers, dates, currency amount, etc., affect the performance of Chinese language processing and the performance of Chinese G2P conversion. The instance in Introduction represents NSWs have a greater propensity than ordinary words to be ambiguous with respect to their interpretation or pronunciation. It is desirable to “normalize” text by replacing NSWs with the contextually appropriate ordinary words. This process is called text normalization. With the corresponding grapheme sequence of NSWs obtained, the corresponding pronunciation can be predicted more accurately. Therefore, text normalization plays an important role in pronunciation prediction of NSWs.

2.2 Chinese Polyphonic Characters

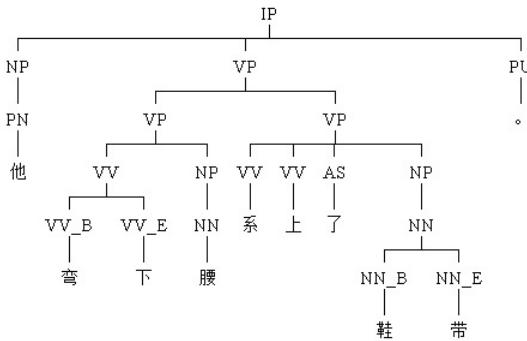
The usefulness of lexical-level cues has been proved in previous works [2,4,6,8]. However, these cues can not resolve all of polyphonic characters, and syntactic and semantic information are also necessary. Therefore, collocation, extracted from syntactic structure, is introduced in these section. The set of Chinese polyphonic characters in “Modern Chinese Dictionary” [1] are investigated referring to “Chinese Polyphonic Character Dictionary” and “Contemporary Chinese Grammatical Knowledge Base”. Because the corresponding pronunciations of interchangeable characters and variant forms of a Chinese character should not be used in Modern Chinese according to [11], these characters turn out to be monophonic characters. Therefore, the number of polyphonic characters decreases to 781. According to the linguistic cues used for disambiguation, these polyphonic characters can be classified into 5 categories as follow:

- Word Form(WF): polyphonic characters act as words or appear in some words.
- POS: each kind of pronunciation corresponds to different POS.
- Named Entity(NE): polyphonic characters with a special pronunciation only are used in named entity.
- Collocation(CO): polyphonic characters can be disambiguated using long span linguistic cues, which can be extracted from syntactic structure.
- Other: the disambiguation needs syntactic and semantic information.

The details are presented in Table 1.

Table 1. Classification of Chinese Polyphonic Characters

Category	Number	Examples
WF	234	“同tòng” only appears in “胡同”; “同tóng” appears in other words
POS	261	“假jiǎ”: adjective or verb; “假jià”: noun
NE	32	“仇qiú”: surname; “仇chóu”: not surname
CO	36	The relation of “系jì” and “鞋带” represents an specific collocation as illustrated in Fig. 1
Other	228	“啊”: a, á, á, à with different emotions

**Fig. 1.** Parsing Tree of sentence “他弯下腰系上了鞋带”

3 Proposed Method

In [12], a character-based syntactic analysis approach integrating morphological and syntactic information is proposed to perform word segmentation, POS tagging and syntactic parsing simultaneously in a unified parsing model, and achieves better performance than the traditional approaches. On this basis, this study further introduces named entity recognition and text normalization into the integrated syntactic parser. These multi-level linguistic cues could be extracted from this new integrated syntactic parser to help pronunciation prediction of NSWs and disambiguation of polyphonic characters. The details will be described in the following sections.

3.1 Pronunciation Prediction of Non-Standard Words

Pronunciation of NSWs can be benefited from the text normalization. As for the text normalization, the traditional rule-based method utilizes a set of rules capturing the characteristics of NSWs in form and context, yet works well only in some specified domains since NSWs are quite different across different domains. Therefore, the combination of rule-based and statistic-based is proposed recently. NSWs are firstly classified by rules or statistical model. If NSWs own only one pronunciation under the category, its pronunciation can be predicted by rules,

otherwise the pronunciation is predicted by statistical model. Consequently, the classification of NSWs is a crucial stage. There are two types of classification: according to the structure of NSWs and according to the meaning of NSWs. The classification according to the structure of NSWs does not determine the pronunciation uniquely, e.g. the NSW “20:30” can be interpreted either time or ratio corresponding to two pronunciations and need to be determined by statistical method.

In this study, NSWs are classified into 14 categories according to the meaning of NSWs and each category can determine the pronunciation uniquely. After classification of NSWs through the unified syntactic parser, a rule-based method implemented with weight finite state transducers(WFSTs), is used to predict the pronunciation. The NSWs classification strategy is shown in Table 2.

Table 2. Non-Standard Words Classification Strategy

Category	Examples	Category	Examples
Cardinal Numbers	100, 1.5吨	Ordinal Numbers	I, (1) , ①
Dates	2013年, 2013/01/01	Times	11时20分, 15:30
Event	“9.11”事件	Postalcode	100871
Phone Numbers	110, 010-65091082	IP Address	192.168.1.1
Money and Currency	¥1000, \$1000	Angle	1° , 1'
Temperature	-1°C	Codes	ISO9001, CA2312
Fraction	1%, 1/3	Ratio	1:2, 1-2

3.2 Disambiguation of Polyphonic Characters

Maximum entropy model can integrate features flexibly and multi-level linguistic cues mentioned above can be used as features. Table 3 and Table 4 represent the basic and compound feature template respectively.

According to the investigation in Section 2, different polyphonic characters should use different features. Therefore, automatic feature selection is introduced, which uses a greedy algorithm and cross validation to get proper feature set for each polyphonic character. If the size of candidate feature set is N and choosing one feature at each iteration, the maximum of iterations is N and the time complexity is $O(N^2)$.

4 Experiments and Results

Our experiments are conducted on the corpus from the People’s Daily of China in 2000, which is randomly divide into two parts: 90% used as the training set of maximum entropy model and 10% as the test set. There are totally 803 polyphonic characters in the corpus. The integrated syntactic parser is trained on Penn Chinese Treebank (CTB 5.0).

Table 3. Basic Feature Template for Maximum Entropy Model

Feature Template	Description
$C_i(-L \leq i \leq L, i \neq 0)$	previous and next characters
$W_i(-L \leq i \leq L)$	current, previous and next words
$POS_i(-L \leq i \leq L)$	POS of current, previous and next words
$SYN_i(-L \leq i \leq L)$	father phrases of current, previous and next words in parsing tree
$PSYN_i(-L \leq i \leq L)$	grandfather phrases of current, previous and next words in parsing tree
$GPSYN_i(-L \leq i \leq L)$	great-grandfather phrases of current, previous and next words in parsing tree
LeftSibling, RightSibling	left and right siblings of current word in parsing tree
PosInWord, PosInSent	relative positions of current character in word and sentence
CO	collocation of current word in parsing tree
Len	length of current word
Punc	punctuation of sentence

L which is the window size of feature, is set 2 in this work.

Table 4. Compound Feature Template for Maximum Entropy Models

Feature Template	Description
$C_iC_{i+1}(-L \leq i \leq L, i \neq 0)$	combination of characters
$W_iW_{i+1}(-L \leq i \leq L, i \neq 0)$	combination of words
$POS_iPOS_{i+1}(-L \leq i \leq L, i \neq 0)$	combination of POS
$SYN_iSYN_{i+1}(-L \leq i \leq L, i \neq 0)$	combination of father phrases

L which is the window size of feature, is set 2 in this work.

The effectiveness of pronunciation prediction of NSWs is investigated firstly. The baseline is a combination of rule-based and statistical method using WFST and N-gram. Results shown in Table 5 represent that the proposed method for pronunciation prediction of NSWs based on integrated syntactic parser improves by 6.90% absolutely and obtains improvement for each category. Because NSWs in categories of IP address, currency and angle do not appear in corpus, there are not their results.

Secondly, to validate the usefulness of multi-level linguistic cues, considering the affect of errors generated by integrated syntactic analysis, maximum entropy models with automatic feature selection are used for all polyphonic characters. The window size of features is 2. The 4-turn cross validation is used in maximum entropy model training process. Traditional maximum frequency guess approach is used as baseline. Results are shown in Table 6. From experimental results, multi-level linguistic cues are useful for disambiguation of polyphonic characters and maximum entropy model can reduce the affect brought by error results of integrated syntactic analysis.

Table 5. Results of NSWs Pronunciation Prediction

Category	Baseline(%)	Proposed Method(%)
Cardinal Numbers	90.02	97.01
Ordinal Numbers	87.43	99.67
Dates	94.73	98.86
Times	96.46	99.63
Event	55.88	100.00
Telephone Numbers	8.00	75.00
Postalcode	0.00	100.00
Codes	35.48	72.72
Temperature	100.00	100.00
Fraction	91.07	99.21
Ratio	73.56	95.68
Total	90.83	97.73

Table 6. Results of Polyphonic Characters Disambiguation

Category	Baseline(%)	ME
Word Form	99.76	99.88
POS	97.34	99.64
Named Entity	95.03	99.04
Collocation	99.01	99.58
Other	90.03	98.16
total	95.64	99.23

5 Conclusions

The performance of Chinese G2P conversion affects the quality of acoustic model in ASR and the intelligibility of speech synthesized by TTS system. This paper investigates two important issues in Chinese G2P conversion – NSWs and Chinese polyphonic characters. For disambiguation of polyphonic characters, the multi-level linguistic cues, including word form, POS, named entity and collocation, are integrated in maximum entropy model to predict the pronunciation. To prevent cascade errors in traditional method for features extraction, an integrated syntactic analysis approach is used to extract these cues together. Besides, the text normalization is integrated in new integrated syntactic analyzer to help predict the pronunciation of NSWs. Experimental results represent that the proposed method are useful for pronunciation prediction of NSWs. And the use of multi-level linguistic cues gets 3.59% improvement in average accuracy rate absolutely than baseline.

Acknowledgment. The work was supported in part by the National Basic Research Program of China (2013CB329304) and National Natural Science Foundation of China (No. 61121002, No.91120001).

References

1. Institute of Linguistics in Chinese Academy of Social Sciences: Modern Chinese Dictionary (the 6th edn.). Commercial Press, BeiJing (2012)
2. Hai Ping, L., Heng, L., Lian Hong, C.: The investigations on Chinese polyphonic characters based on context. In: 4th National Symposium on Human-Machine Speech Communication, pp. 231–234, BeiJing (1996)
3. Min, Z., LianHong, C.: A new rule-based method of automatic phonic notation on polyphones. In: The Second National Symposium on Computational Linguistics, pp. 238–243 (2004)
4. Wern-Jun, W., Shaw-Hwa, H., Sin-Horng, C.: The broad study of homograph disambiguity for Mandarin speech synthesis. In: Proceedings of International Conference on Spoken Language Processing, pp. 1389–1392 (1996)
5. Jie, W., E.D., X., Zhiyong, L., Rou, S.: The research of automatic phonetic notation on Chinese polyphonic words. In: JSCL, pp. 534–539 (2005)
6. Zirong, Z., Min, C.: A statistical approach for grapheme-to-phoneme conversion in Chinese. Journal of Chinese Information Processing 16, 39–45 (2002)
7. Fangzhou, L., Qin, S., Jianhua, T.: Maximum entropy based homograph disambiguation. In: National Conference on Man-Machine Speech Communication, pp. 19–24 (2007)
8. Hao, T., X.J., L., X.H., W., H.S., C.: A Combination of Rule-based and Statistical-based methods for Notation of polyphonic characters. In: National Conference on Man-Machine Speech Communication, pp. 508–511 (2005)
9. F.L., H.: Disambiguating effectively Chinese polyphonic ambiguity based on unify approach. In: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics and Cybernetics, Kunming, pp. 3242–3246 (2008)
10. Fangzhou, L., You, Z.: Polyphone disambiguation based on tree-guided TBL computer engineering and applications. Computer Engineering and Applications 12, 137–140 (2011)
11. Ministry of P. R. China: Law of the People's Republic of China on the Standard Spoken and Written Chinese Language. Standing Committee of the 9th National People's Congress (2001)
12. Wu, X., Zhang, M., Lin, X.: Parsing-based Chinese word segmentation integrating morphological and syntactic information. In: 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 114–121. IEEE (2011)

Blind Quality Assessment on Binary Seal Images

Chengyun Wang, Zhanlong Hao, and Youbin Chen

Graduate School at Shenzhen, Tsinghua University, China

{wang.chengyun11,hao.zhanlong10}@alum.sz.tsinghua.edu.cn,
chenyb@sz.tsinghua.edu.cn

Abstract. In this paper, we propose a method to segment seals and evaluate their quality. Seals with inferior qualities are not suitable for verification. To enhance the robustness of seal system, we put forward a strategy to assess the quality of extracted seal images. First we propose a method to segment seals and get the characters. Then by human assessment, we assign different characters with proper scores as ground-truth. We utilize a series of features and the SVR regression to predict the quality. Finally we use Optical Character Recognition rates to test the effectiveness. Experimental results prove that our proposed method is very effective.

Keywords: seal segmentation, image quality, blind assessment, regression.

1 Introduction

Seals are widely used in oriental countries and a mass of seal images need to be verified every day, hence automatic seal systems [1] are of huge commercial value. Intact seal processing systems typically focus on a variety of functions (seal recognition, retrieval, and verification) [2]. However, for most systems the first decisive procedure is to extract the seal imprint. Extraction is very important because it influences the accuracy largely. The quality of seal images varies much (Fig. 1 (a) is a low quality sample with some characters filled and this factor always leads failure in identification systems. Fig. 1 (b) shows a high quality sample compared with Fig. 1 (a)) and how to measure the quality is of great importance. Therefore this paper pays attention to the blind quality assessment of extracted seal imprints.

Hitherto many methods for document image quality metrics have been proposed. In [3] and [4], quality measures (white speckle, small speckle, broken character, etc) are utilized to predict Optical Character Recognition (OCR) accuracy. In [5], an evaluation method to improve character recognition rates is introduced. Gray distribution feature and neural network classifier are applied to assess Chinese characters. [6] presents a dual framework (a degradation classifier and two regressions) to predict human perception of English characters. Features based on morphological operations and filters are described to train the Multi-layer Perceptron (MLP). In [7] a measure for binary document images via human visual is introduced and the distance of pixels between samples and the reference images is utilized. [8] proposes a method to evaluate binarization results. Broken line structures, noise in homogeneous areas are taken into account.

This paper is organized as follows: Section 2 introduces a method to segment seals and to get the characters in seal imprints. Section 3 shows how to evaluate these characters. Section 4 is the experimental results. Section 5 is the summary of this paper.

2 Character Segmentation

2.1 Mathematical Transformation

The binary seals are shown in Fig. 1. Because all these seals are from real bank items, we remove some characters by gray spots for security reason. Traditional methods of segmenting circular seals mainly use mathematical transformation to convert the circle into a rectangle region [2], [9], and [10]. There is a gap between the frame and main character region. It is easy to get the characters through projection as showed in Fig. 1 (c).

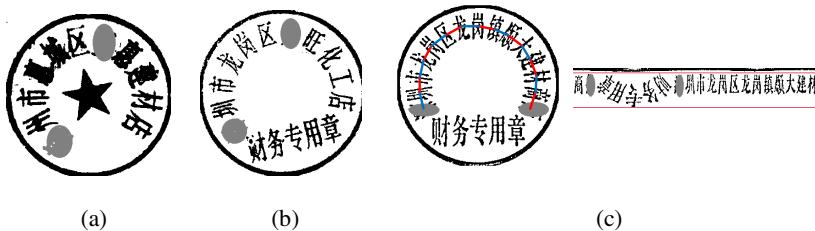


Fig. 1. (a) A low quality extracted seal sample with some characters filled that is not suitable for seal systems. (b) A good seal sample compared with (a). (c) The transform of circular seal.

2.2 Auxiliary Region Removal

The next step is to wipe off the auxiliary part (“财务专用章”) to get the characters we need. In [2] by vertical projection, the regions whose width is over the median are assumed as auxiliary characters. However, this method often fails when characters are adhesive. To overcome the shortcomings, we use template matching algorithm. In China, circular seals are mainly used in some formal occasions. The characters in auxiliary region are always alike (“财务专用章”) that is claimed by laws. By this observation, we exploit templates to delete it. Fig. 2 (a) and (b) show two templates we use in the experiment. A slide window is used to search the best location that maximizes $R(x, y)$.

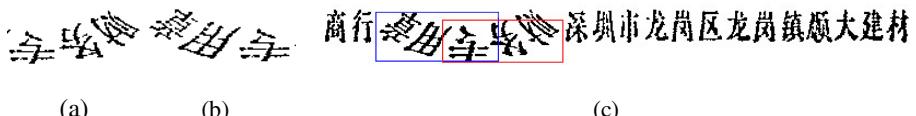


Fig. 2. (a) The template image for locating the start points. (b) The template image for locating the end points. (c) The result of template matching process.

$$R(x,y) = \sum_{x',y'} (T(x',y') * I(x+x',y+y')). \quad (1)$$

Where $T(x,y)$ presents the template image and $I(x,y)$ is the rectangle region. $R(x,y)$ denotes the result of convolution.

2.3 Character Segmentation

For the character line we obtain, what we need to do next is to segment them into single characters. Horizontal projection segmentation applied by [2] and [9] assumes that spaces between adjacent characters are obvious without noise. However, this straightforward method does not perform very stable.

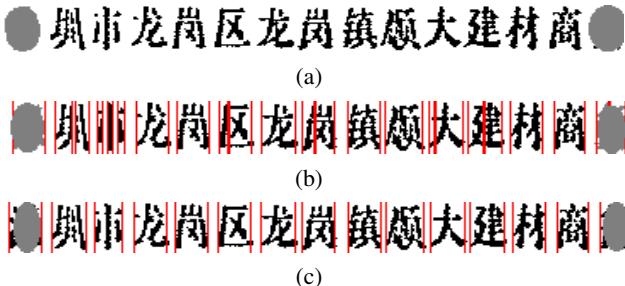


Fig. 3. (a) Original image. (b) Over-segmentation. (c) Merged by path search method

Dissection and search methods [11] are two traditional approaches in segmentation. In [12], these two methods are combined. Firstly the dissection method is applied to over-segment characters. Secondly different optimization goals (in [12], recognition confidence is the metrics to select final boundaries) are set to search for the optimal segmentation. It is obvious that distances of the centers of adjacent characters (the red or blue lines shown in Fig. 1 (c)) are almost the same. First of all, we use normal methods to get all characters over segmented as showed in Fig. 3 (b). Then we merge some boundaries into a new one. With a view to the uniform distances between adjacent characters, of all possible segmentations we search for a segmenting path that minimizes *Grade* defined in formula (2). The one that minimizes *Grade* indicates that characters are cut as uniform as possible.

$$Grade = \frac{\sigma}{\mu}. \quad (2)$$

$$\mu = \frac{1}{N} \sum_{i=1}^N c_i, \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (c_i - \mu)^2}, c_i = x_{i+1} - x_i, c_i < c_{max} (i = 1, 2, \dots, N). \quad (3)$$

Where x_i is the coordinate of the center of the bounding box, c_i is the center distance between adjacent bounding boxes, μ is the mean width of all bounding boxes, σ is the

variance of all boundaries, and N is the number of bounding boxes. To avoid merging all characters into one, we set an upper limit c_{max} (c_{max} is two times as the media width of all bounding boxes in over-segment step). Fig. 3 (c) shows the final result.

3 Character Quality Assessment

Real-world seals suffer from much degeneration (seals carelessly imprinted always contain broken or filled strokes). There are many studies to reduce the damages brought by these. However, there is not an existing method that could work well in all cases. So far there have been few researches about evaluating the quality of seal images. Via evaluation, we could pick up seals inappropriate to deal with.

3.1 Ground Score by Human Visual Assessment

First of all, we use segment algorithm above to collect 542 Chinese characters. All these characters are normalized to the same size 60*38. Some character strokes are sticky or filled, while some are broken as showed in Fig. 4 (a). To employ human assessment to train our system, we had 20 volunteers from Tsinghua University to sort the data into three categories (good, moderate, and bad). Then we use 1, 2, 3 scores to represent bad, moderate, and good quality respectively. The mean scores of all people for each character serve as the ground-truth score. The score 1 presents the worst quality while score 3 means the best quality.

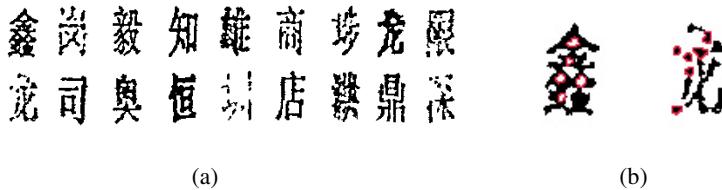


Fig. 4. (a) Chinese characters segmented from seals with different qualities. (b) A character with small holes and a character with broken strokes.

3.2 Features and the SVR Predictor

[6] studies the inner relations between human visual assessment and degeneration levels of English typewritten characters. It uses 21 features to measure the faded degree from various aspects. All these features could be classified into three categories: morphological-based features, noise-removal-based features, and some spatial characteristic features. Through erosion, dilation, operations, and a series of filters, it can imply the stroke width or the smooth of the edge rough. The details can be found in [6]. As the deterioration in Chinese seals shares many similarities with it (many seal characters are broken or filled for careless stamping), we take the same features for reference to assess Chinese characters.

Chinese characters are different from English words (the number of Chinese characters is much larger) and are more complex. In order to describe faded words more comprehensively, we add two typical features: the histogram of the number of structures in different sizes and the histogram of the number of holes in different sizes. For broken characters, there exist some small connected components. Filled characters always have small holes leaded by abnormal strokes as showed in Fig. 4 (b).

As we know, SVR is widely used for regression. By SVR predictors, we could find proper weights to combine all features and map them into scores. With the 23 features, we make use of RBF kernel to train our models. By this, we get the predicted scores of all characters. Then we use their average score to represent the quality of whole seal.

$$S_{seal} = \frac{1}{N} \sum_{i=1}^N S_i, (i = 1, 2, \dots, N). \quad (4)$$

Where S_i is the predicted score of a single character, N is the number of the characters and S_{seal} is the final score of a seal.

4 Experimental Results

4.1 The Result of Segmentation

The performance of the proposed segmentation method was evaluated using 550 seal images. Some are of poor quality and this brings much difficulty in segmentation. Among all images, 486 were accepted by our systems. We successfully segmented 477 and got accuracy of 98.1% as showed in Table 1.

Table 1. Result of seal character segmentation

Method	Acceptance rate	Correct rate
Method in [2]	57.82%(318/550)	77.7%(247/318)
Proposed method	88.36%(486/550)	98.1%(477/486)

Our proposed method shows its advantage. The merging strategy used in traditional method is too simple (when two boundaries are close to each other, it merges them into one). Based on assumption of uniform distances between adjacent machine-printed characters, the *Grade* is defined. Because the geometrical structure of a seal is taken into account, it is not only accurate but also stable. The result shows that our method is better than traditional methods. However, for some seals the size of its auxiliary region is very abnormal. The template matching step may fail in this case.

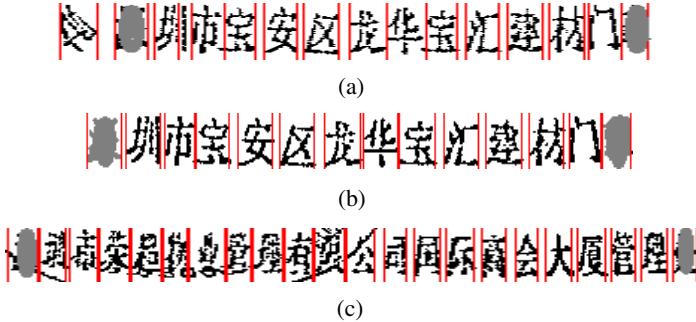


Fig. 5. (a) Method by [2]. (b) Proposed method. (c) The result of proposed method (a sample of low quality and refused by method [2]).

4.2 The Result of Quality Assessment

By segmentation, we collect about 7000 Chinese characters. It is time-consuming to do the evaluation by humans, so we just select 542 characters for our assessment experiments. 292 were used for the train set and the left 250 were regarded as the test set. Another test set in the last row of Table 3 is a set containing 6690 characters. Two indexes were adopted to evaluate the performance: the Pearson Correlation Coefficient (PCC) and the Relative Absolute Error (RAE) in [6].

$$RAE = \frac{1}{n} \sum_{k=1}^n \frac{|S_k - \hat{S}_k|}{S_k}. \quad (5)$$

Where S_k is the ground-truth score, \hat{S}_k is the predicted score and n is the number of characters. Fig. 6 (a) shows the linear relations between predict scores and the ground-truth scores. PCC fluctuates much when the number of people is small. As the number increases, PCC becomes stable gradually as showed in Fig. 6 (b).

Table 2. Result of quality assessment

Data set	RAE	PCC
Train set	0.1271	0.9216
Test set	0.1666	0.8177

In seal retrieval systems, the target seal are found by recognition. Therefore the recognition rate determines the reliability of a system. To prove the effectiveness of this strategy, we use an OCR engine to recognize all characters. The advantages of our assessment methods can be seen from Table 3. Characters with predicted scores above 2 share a higher recognition rate than those below 2. It is also higher than average recognition rate. Thus we could use our assessment method to discriminate the good-quality seals from the poor-quality ones.

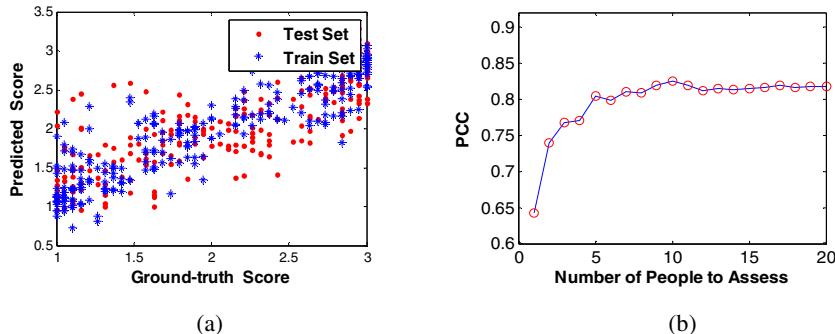


Fig. 6. (a) Predicted scores by systems and the ground-truth scores. (b) PCC varies with different number of people to evaluate.

Table 3. OCR accuracy

Data set	Recognition rate (score>2)	Recognition rate (score≤2)	Average Recognition rate
Train set	85.64%	37.23%	67.46%
Test set	76.26%	26.2%	54.96%
Another test set	87.48%	52.42%	83.26%

5 Summary

In this paper, we propose a new method to segment circular seals. Dissection and search methods used in document segmentation are adopted in the segmentation step. To enhance the robustness of seal systems, we first put forward a blind quality assessment method to evaluate the extracted seals. Through human vision assessing Chinese characters, we could predict the quality of seals. The OCR results prove the effectiveness of this framework. In the future, we will assess the quality of color seal imprints before they are extracted from document images.

References

1. Ueda, K., Matsuo, K.: Automatic seal imprint verification system for bank check processing. In: Third International Conference on Information Technology and Applications, ICITA 2005, vol. 1, pp. 768–771 (2005)
2. Ren, C., Liu, D., Chen, Y.: A new method on Segmentation and Recognition of Chinese Characters for Automatic Chinese Seal Imprint Retrieval. In: International Conference on Document Analysis and Recognition, ICDAR 2011, pp. 972–976 (2011)
3. Blasndo, L.R.: Prediction of OCR accuracy using simple image features. In: International Conference on Document Analysis and Recognition, ICDAR 1995, vol. 1, pp. 319–322 (1995)

4. Cannon, M.: Quality assessment and restoration of typewritten document images. *International Journal on Document Analysis and Recognition, IJDAR* 1999 2(2-3), 80–89 (1999)
5. Liu, C.: Degraded Character Recognition by Image Quality Evaluation. In: *International Conference on Pattern Recognition, ICPR* 2010, pp. 1908–1911 (2010)
6. Obafemi-Ajayi, T.: Character-based Automated Human Perception Quality Assessment in Document images. *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans* 42(3), 584–595 (2012)
7. Lu, H.: An objective distortion measure for binary document images based on human visual perception. In: *International Conference on Pattern Recognition, ICPR*, vol. 4, pp. 239–242 (2002)
8. Trier, O.D.: Evaluation of binarization methods for document images. *IEEE, Trans. Pattern Analysis and Machine Intelligence, PAMI* 17(3), 312–315 (1995)
9. Ren, C., Chen, Y.: A New Method for Character Segmentation and Skew Correction on Chinese Seal Images. In: Qian, Z., Zhihong, Q., Cao, L., Su, W., Wang, T., Yang, H. (eds.) *Recent Advances in CSIE. LNEE*, vol. 128, pp. 263–268. Springer, Heidelberg (2012)
10. Liu, H.: Automatical seal Image retrieval method by using shape feature of Chinese characters. In: *International Conference on Systems, Man and Cybernetics*, pp. 2871–2876 (2007)
11. Casey, R.G.: Strategies in character segmentation: a survey. In: *International Conference on Document Analysis and Recognition, ICDAR*, vol. 2, pp. 1028–1033 (1995)
12. Fujisawa, H.: Segmentation Methods for character Recognition: from segmentation to document structure analysis. *Proceedings of the IEEE* 80(7), 1079–1092 (1992)

An Improved 3D Edge Surface Tracking Algorithm Based on 3D Fractional-Order Differentiation within Confocal Microscopy Images

Yu Ma¹, Yanning Zhang², and Lisheng Wang³

¹ Institute of Computer Science, Northwestern Polytechnical University , Xi'an , 710129, China

² Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University,
Shanghai, 200240, China

mayu.ningxia@gmail.com

Abstract. Fractional-order differentiation enhances the image nonlinearly, but only has been applied in the 2D image. The 2D fractional differentiation operator is extended to 3D and the 3D fractional differentiation discrete filtering masks are deduced. The 3D fractional differentiation is implemented to improve the traditional 3D image edge surface tracking algorithm in two aspects. Firstly, the 3D data fields of neuron slices are enhanced by the 3D fractional differentiation, ensures more detailed structures of edge surfaces with low contrast are extracted. Then integral-order gradient is modified by the fractional differentiation to get more 3D detailed structures. The proposed method has been applied to 3D confocal microscopy images, and more 3D detail structures of neuron are tracked compared to the traditional 3D edge surface tracking algorithm.

Keywords: Fractional-order differentiation, Image enhancement, Confocal microscopy image, 3D edge surface tracking algorithm, Neuron.

1 Introduction

The confocal microscopy images are mainly applied to the research of neuroanatomy and the auxiliary diagnosis of nervous system disease [1][2], the 2D image cannot describe the function of the neuron accurately for its complex structures. The major structure of a neuron is the nerve cell body, and the dendrite is attached to the cell body and extended from thick to thin, the shape is like a tree, the extended terminal is the dendritic spine. The number of dendrite' branches and the area of dendritic spine determine the simulation areas of the neuron. The high accuracy 3D confocal microscopy images can depict the cell body, the dendrite and the dendritic spine clearly, which possess the greatest reference value of the research on brain function. But lower image contrast brings a big difficulty for reconstructing the high accuracy of the 3D structures of the subtle dendritic spine. Significant progress has been obtained in the research on 3D visualization with volumetric data, as a result, the visualization equipments can be extended to the PC from the special graphics and image station, and the speed of the visualization and the amount of researchers are increasing

rapidly. A great deal of research achievements [3] [4] [5] [6] have obtained in the 3D visualization of confocal microscopy images. At the present there are two kinds of methods in 3D reconstruction: volume rendering and surface rendering. The method based on gradient and Laplacian operator has overcome the shortcoming of the Marching Cubes method [7] [8], and can extract 3D edge surfaces with high accuracy, that is subvoxel accuracy, but the improper setting of gradient magnitude threshold leads to a negative influence, the big threshold generates an inadequate extraction, on the contrary, the small one brings the noisy edge surfaces of background and overwhelms the foreground objects. A tracking 3D edge surface extraction algorithm based on the coplanar principle is put forward to overcome the disadvantages of the above method, and it shows the insensitivity of the gradient magnitude threshold [9] [10]. The tracking algorithm has been used to extract 3D edge surface of confocal microscopy images, and obtain good results. However, the simple cell body has higher contrast, while the complex dendritic spine has lower contrast, thus the improved tracking algorithm is unable to extract more detailed 3D structure of the dendritic spine. In order to extract more detailed structure of the dendritic spine, in this paper, the fractional differentiation [11] [12] has been implemented to improve the tracking algorithm due to the enhancement characteristic.

Fractional order differentiation (FD) is becoming an important branch of mathematic analysis, which is the extension of the integral order differentiation. FD can not only enhance the high frequency nonlinearly, but also preserve the medium frequency to a certain extent, meanwhile, can also keep the low frequency and DC component nonlinearly of a signal [13]. Yifei Pu firstly introduces the fractional differentiation to digital image processing and applies it to enhance the 2D image, which is helpful to detect more detailed information of a 2D image [14]. A great deal of research achievements has been obtained [15] [16] [17]. This paper extends 2D-FD to 3D field, here we call 3D-FD, and improves the 3D edge surface tracking algorithm to track more detailed structures within 3D images. In the proposed method, in order to enhance the low contrast structure, 3D volumetric data is first enhanced by the 3D-FD, furthermore, the fractional gradient instead of integral gradient for tracking the detail information more accurately.

The paper is organized as follows: The first part introduces the 3D edge surface tracking algorithm. On the basis of introducing 2D-FD, the second part extends 2D-FD to 3D (3D-FD), and deducts the discrete filtering masks of 3D-FD. The third section suggests the improved tracking algorithm based on 3D-FD. The fourth part discusses the experimental results, and the final part concludes the paper.

2 The 3D Edge Surface Tracking Algorithm

Based on Laplacian operator and the gradient magnitude threshold, the 3D edge surface tracking algorithm [9] makes use of the coplanar principle of the adjacent cubes in 3D regular dataset and extracts the 3D subvoxel edge surfaces. It has been used successfully to extract and track the high accuracy 3D edge surfaces within CT, MRI, confocal microscopy and other biomedical images. The detailed information about the 3D edge surface tracking algorithm has been proposed in [8] and [9].

3 The Fractional-Order Differentiation

The fractional differentiation [12] has been applied to engineering and technical field widely in recent years. Yifei Pu et al. have successfully implemented the method into signal processing and image processing [14] [15], furthermore, proposed filter masks and numerical calculation of fractional differentiation, and obtained lots of research achievement. In this paper, we extend 2D fractional-order differentiation to 3D field and deduce the 3D filtering masks, which used to improve the 3D edge surface tracking algorithm.

3.1 The 2D Fractional-Order Differentiation

In the continuous function, the fractional-order differentiation [11] is the fractional-order extension of the integral one. For any given square integrable energy function, the fractional-order differentiation can be defined as follows:

$$D^v f(t) = \frac{d^v f(t)}{dt^v} \quad (1)$$

Its Fourier transformation is:

$$D^v f(\omega) = (i\omega)^v \bullet f(\omega) (v \in R^+) \quad (2)$$

Under the Euclid distance measure, fractional-order differentiation has three definitions: Grünwald-Letnikov, Riemann-Liouville and Caputo definitions, the Grünwald-Letnikov ($G-L$) definition has been applied widely. $\forall v \in R$, if the signal $f(t) \in [a, t] (a < t, a \in R, t \in R)$ exists continuous differentiation with the order $m+1$, when $v > 0$ and v is non-integer, the v -order fractional-order differentiation of $G-L$ definition is shown as below [11]:

$$D^v f(t) = \lim_{h \rightarrow 0} h^{-v} \sum_{r=0}^n \left[\frac{-v}{r} \right] f(t - rh), \quad (3)$$

where $\left[\frac{-v}{r} \right] = \frac{(-v)(-v+1)\dots(-v+r-1)}{r!}$. More detailed descriptions of the fractional differentiation and the derivation of 2D discrete filter masks can be seen in the references [12] [13] [14].

3.2 The 3D Fractional-Order Differentiation

The extension from the 2D fractional differentiation to 3D is similar to the extension from one dimensional to 2D, and the 3D spatial information should be considered. We can deduce the continuous 3D fractional-order differentiation $\frac{\partial^v f(x, y, z)}{\partial x^v}$ of the function $f(x, y, z)$, taking the negative direction on x axis for example:

$$\begin{aligned} \frac{\partial^v f(x, y, z)}{\partial x^v} \approx & f(x, y, z) + (-v)f(x-1, y, z) + \frac{(-v)(-v+1)}{2}f(x-2, y, z) \\ & + \frac{(-v)(-v+1)(-v+2)}{6}f(x-3, y, z) + \dots \frac{\Gamma(n-v-1)}{(n-1)!\Gamma(-v)}f(x-n+1, y, z) \end{aligned} \quad (4)$$

3.3 The Discrete Filtering Masks of 3D Fractional Differentiation

According to the discrete filtering mask of the 2D fractional-order differential above [13] [14], we can also easily extend the 2D case to 3D by considering the 3D spatial information. In 3D slices stack image, there are total eighteen-direction filtering masks in 3D space, three layers of the slices are depicted: the layer $(z-1)$, (z) and $(z+1)$. The voxel X and each of its neighbors X_i ($i=1\dots18$) constitute the direction $\langle X, N_i \rangle$, eighteen neighbors determined eighteen directions [18]. Similar to eight directions corresponds to eight filter masks in 2D field, and eighteen directions determine eighteen filtering masks in 3D images. The 3D filter masks can be described as follows:

$$V_n(f) = \sum_{m=1}^4 F_m(f) * \Psi(m) \quad (7)$$

where $V_n(f)$ is the result of the n -direction ($n=1, 2\dots18$) filtering mask, $F_m(f)$ represents the gray value of the m neighbor of the voxel X , $\psi(m)$ is the m coefficient in each mask. In order to reduce the computation due to the large volumes of 3D volumetric data, here we take $m=4$, and then there are total seventy two filtering neighborhood voxels in eighteen directions. The coefficient of the 3D masks $\psi(m)$ is similar to the 2D, expressed as follows:

$$\Psi(m) = \frac{m^{-v} - (m-2)^{-v}}{\Gamma(-v)(-2v)} \quad (8)$$

According to the formula (7) and (8), the enhanced result $D(f)$ of the voxel X by using the fractional differentiation can be described below:

$$D(f) = \sum_{n=1}^{18} V_n(f) * \frac{V_n(f)}{\sum_{\varepsilon=1}^{18} V_\varepsilon(i, j, k)} \quad (9)$$

Given that the scale of 3D volumetric data is $M \cdot M \cdot L$, filter each voxel by using formula (9) until traverse all the $M \cdot M \cdot L$ voxels, and obtains the enhanced 3D volumetric data of the images.

4 The Improved Edge Surface Tracking Algorithm Based on 3D Fractional-Order Differentiation

The fractional differentiation is able to enhance the edge feature, sharpen the image and clarify the details. By adjusting the fractional order, the enhancement effect can be controlled, furthermore, the computation accuracy of 3D fractional-order differentiation can be adjusted by the size of the filtering mask [15] [16]. Firstly, we apply the extension result 3D fractional-order differentiation to enhance the 3D volumetric data, it is helpful to track more detailed structure of dendritic spine in neurons. The second improvement is replacing the integral gradient with the fractional gradient in the tracking process and improves the tracking accuracy.

The improved tracking algorithm is as follows:

(1) Detect edge cubes by using formula (2), choose the integral gradient magnitude T_{high} as the detection threshold to ensure the detected edge cubes to stand for each part of the 3D structure.

(2) Choose the seed cubes from the detected edge cubes in step 1.

(3) Enhance and sharpen the 3D volumetric data by using 3D-FD proposed in section 2, track the detailed structure by using the fractional gradient magnitude threshold T_{low} , the mathematical constraint conditions can be described below:

$$\begin{cases} \nabla^2 f'(x, y, z) = 0 \\ \|\nabla^v f(x, y, z)\| \geq T_{low} \end{cases} \quad (10)$$

where $\nabla^2 f'(x, y, z)$ represents the Laplacian value of the 3D image $f'(x, y, z)$, which has been enhanced by 3D-FD and is different from $\nabla^2 f(x, y, z)$ in the formula (2). $\nabla^v f(x, y, z)$ ($0 < v < 1$) is a fractional-order gradient with the order v , it is also different from the integral gradient $\nabla f(x, y, z)$ in formula (2).

(4) According to the coplanar principle of the adjacent cubes, employ the 3D region growing method, use the dynamic chain stacks and the depth first search (DFS) method to track the rest edge cubes, which are corresponding to the other detail structures in the edge cubes restrained in formula (10).

(5) Calculate the polygon in each edge cubes, obtain the triangular plates of the polygons, and then visualize the triangular plates by using the technology of computer graphics, that is, visualize the tracked edge surfaces.

5 Experimental Results

The reconstructions of the neuron confocal microscopy image with the volumetric data scale 180·206·35 are carried out by the tracking algorithm and the improved algorithm based on 3D fractional-order differentiation, then the two tracked results are compared and analyzed.

5.1 The Edge Surface Tracking Algorithm

The edge surface tracking algorithm has been proposed to overcome the shortcomings of the 3D edge surface extraction algorithm. The tracking algorithm first extracts most of the edge cubes with a higher integral gradient magnitude threshold T_{high} , ensures to reduce the noisy cubes. Then choose some edge cubes as seed cubes, on the basis of the coplanar principle of adjacent cubes in 3D data field, track the other edge cubes of detailed structures with a small integral gradient magnitude threshold T_{low} . That is, a relatively smaller tracking threshold T_{low} can ensure track more dendritic spines of the neuron [9]. Fig.1(a) indicates the tracked result by the edge surface tracking algorithm, with the detecting threshold $T_{high} = 3000$ and the tracking threshold $T_{low} = 50$. By the comparison of the edge surface extraction algorithm, the tracking algorithm has improved the extraction algorithm greatly [9].

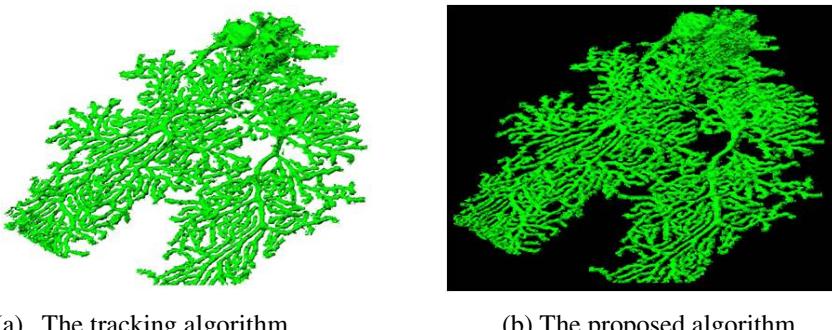


Fig. 1. The tracking algorithm and the proposed algorithm

5.2 The Improved 3D Edge Surface Tracking Algorithm Based on Fractional-Order Differentiation

The tracking algorithm can track more 3D detailed structures than the extraction algorithm, it also overcomes the defect of the extraction method, and however, the tracked results of the dendrites and dendritic spines are still not enough. Thus the improved tracking algorithm based on 3D fractional-order differentiation has been proposed in this paper, and the experimental results are shown in Fig.1 (b). Fig.1 (b) describes the results of the improved tracking algorithm, with the order 0.75. It is obviously that the improved method proposed in this paper tracks more detailed information of the dendrite and the dendritic spine than the tracking algorithm.

The more in-depth comparison between the tracking algorithm and the improved algorithm has been explored, and the local region of the results of the two methods is depicted in Fig.2 (a) and (b) respectively. It can be easily seen that the improved algorithm tracked more detailed structure of dendrites and dendritic spines.

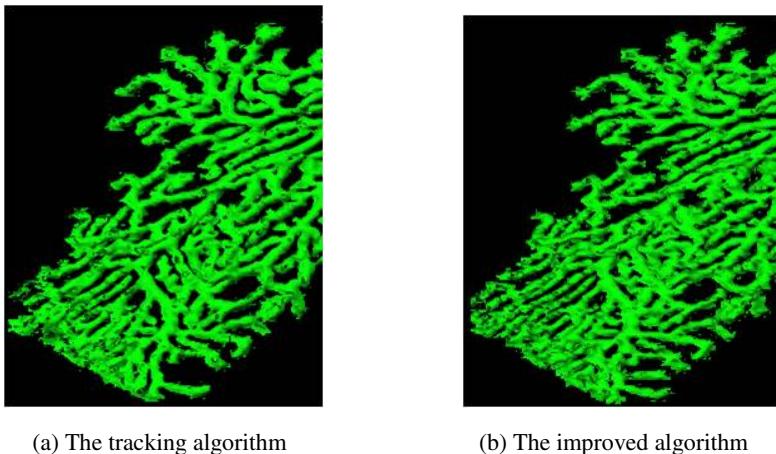


Fig. 2. The local region comparison between the tracking algorithm with the improved algorithm based on fractional differentiation

6 Conclusions

In this paper, the 2D fractional-order differentiation is analyzed and has been extended to the 3D field, and the 3D discrete filtering masks have been deduced. Since the enhancement feature of the fractional differentiation, the 3D fractional-order differentiation has been applied to improve the 3D edge surface tracking algorithm and has tracked more 3D detailed structure of neuron within the confocal microscopy images. The experimental results and the comparison demonstrate that the improved tracking algorithm can track more abundant 3D structures of dendrites and dendritic spines with lower contrast region. The proposed 3D fractional differentiation method also can be applied to other research fields within 3D biomedical images.

References

1. Dima, A., Scholz, M., Obermayer, K.: Automatic segmentation and skeletonization of neurons from confocal microscopy images based on the 3-D wavelet transform. *IEEE Transactions on Image Processing* 11(7), 790–801 (2002)
2. Konstantinidis, I., Santamaría-Pang, A., Kakadiaris, I.A.: Frames-based denoising in 3D confocal microscopy imaging. In: 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2005. IEEE (2005)
3. Al-Kofahi, K.A., et al.: Rapid automated three-dimensional tracing of neurons from confocal image stacks. *IEEE Transactions on Information Technology in Biomedicine* 6(2), 171–187 (2002)
4. Bucher, D., et al.: Correction methods for three-dimensional reconstructions from confocal images: I. Tissue shrinking and axial scaling. *Journal of neuroscience methods* 100(1), 135–143 (2000)

5. Cai, H., et al.: Repulsive force based snake model to segment and track neuronal axons in 3D microscopy image stacks. *NeuroImage* 32(4), 1608 (2006)
6. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics* 21(4) (1987)
7. Wang, L., et al.: Template-matching approach to edge detection of volume data. In: *Proceedings of the International Workshop on Medical Imaging and Augmented Reality*. IEEE (2001)
8. Wang, L., et al.: A computational framework for approximating boundary surfaces in 3-D biomedical images. *IEEE Transactions on Information Technology in Biomedicine* 11(6), 668–682 (2007)
9. Yu, M., et al.: A novel algorithm for tracking step like edge surfaces within 3D images. *Journal of Computer-Aided Design&Computer Graphics* 19(3), 329–333 (2007)
10. Yu, M., et al.: An automatic surface extraction for volume visualization. In: *2011 Third International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, vol. 1. IEEE (2011)
11. Oldham, K.B., Spanier, J.: The fractional calculus, vol. 17 (1974)
12. Tenreiro Machado, J.A.: Analysis and design of fractional-order digital control systems. *Systems Analysis Modelling Simulation* 27(2-3), 107–122 (1997)
13. Pu, Y.F., et al.: Fractional differential approach to detecting textural features of digital image and its fractional differential filter implementation. *Science in China Series F: Information Sciences* 51(9), 1319–1339 (2008)
14. Yi-Fei, P.: Application of fractional differential approach to digital image processing. *Journal of Sichuan University (Engineering Science Edition)* 3, 022 (2007)
15. Zhuzhong, Y., et al.: Image Enhancement Based on Fractional Differentiations. *Journal of Computer Aided Design & Computer Graphics* 20(3), 343–348 (2008)
16. Pu, Y.-F., Zhou, J.-L., Yuan, X.: Fractional differential mask: a fractional differential-based approach for multiscale texture enhancement. *IEEE Transactions on Image Processing* 19(2), 491–511 (2010)
17. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Image enhancement and denoising by complex diffusion processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(8), 1020–1036 (2004)
18. Zesheng, T.: *The visualization of three dimensional data fields*, vol. 12. Tsinghua University Press (1999)

A Multiobjective Fuzzy Clustering Algorithm Based on Robust Local Spatial Information for Image Segmentation^{*}

Feng Zhao¹, Hanqiang Liu², and Jiulun Fan¹

¹ School of Telecommunication and Information Engineering,
Xi'an University of Posts and Telecommunications, Xi'an, P.R. China
fzhao.xupt@gmail.com, jiulunf@163.com

² School of Computer Science, Shaanxi Normal University, Xi'an, P.R. China
max.hqliu@gmail.com

Abstract. To obtain the satisfying performance of noisy image segmentation, a multiobjective fuzzy clustering algorithm based on robust local spatial information (MFC_RLS) is proposed. In this method, the robust local spatial information derived from the image is introduced into fitness functions which utilize the fuzzy compactness and fuzzy separation among the clusters. In addition, after producing the set of non-dominated solutions, the final segmentation result is chosen by a validity index with the robust local spatial information. Experimental results show that MFC_RLS behaves well in segmenting noisy images.

Keywords: Image segmentation, Multiobjective fuzzy clustering, Cluster validity measure, Robust local spatial information.

1 Introduction

Image segmentation is one of the most difficult tasks in image analysis. It aims to divide an image into several non-overlapping meaningful regions with homogeneous characteristics [1]. Generally, image segmentation is basically clustering of the pixels in the image according to some criteria. Fuzzy c-means (FCM) clustering algorithm is the most popular clustering algorithm and have been successfully applied into image segmentation. However, FCM is very sensitive to noise, outliers and other imaging artifacts due to not considering any spatial information in image context. In order to overcome these problems, many researchers modified the objective functions of FCM [2-6] by introducing the spatial information derived from the image. It is well known

* This work is supported by the National Natural Science Foundation of China (Grant Nos. 61102095 and 61202153), the Natural Science Basic Research Plan in Shaanxi Province of China (Grant No. 2012JQ8045), the Scientific Research Program Funded by Shaanxi Provincial Education Department (Grant No. 11JK1008), and the Research Fund Program of Key Lab of Intelligent Perception and Image Understanding of Ministry of Education of China (Grant No. IPIU012011008).

that the performance of clustering algorithms relies on the cluster criteria or objective function. Both traditional FCM and FCM algorithms with spatial information only consider the total distance within the clusters and utilize the single objective function for data clustering. These methods tend to be very effective for compact, spherical or well-separated clusters, but they may fail to cluster more complicated data. Moreover, these methods are quite sensitive to the initialization of cluster center or membership function and may get into the local optimization.

Recently, many researchers attempted to optimize several cluster validity measures and proposed many clustering approaches based on multiobjective optimization [7,8]. Handl and Knowles [7] proposed a multiobjective evolutionary clustering algorithm in which deviation and connectivity were used as two objective functions to be optimized simultaneously. In the multiobjective evolutionary clustering algorithms [7,8], the cluster validity measures almost are J_m [9], Xie–Beni (XB) [10] and PBM index [11] which are used to optimize and integrated the non-dominated solutions set. In the existing multiobjective clustering algorithms, most of the validity measures can not consider the spatial information of the image. Therefore, when the image is heavily contaminated by noise, these methods may not obtain the satisfying segmentation results.

In this paper, a multiobjective fuzzy clustering algorithm based on robust local spatial information (MFC_RLS) for image segmentation is proposed. Firstly, the robust local spatial information is introduced into the fitness functions which utilize the fuzzy compactness and fuzzy separation among the clusters. In MFC_RLS, we adopt Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [12] as the underlying optimization strategy. After producing the set of non-dominated solutions by optimizing the above fitness functions, the final result is chosen by a validity index with the robust local spatial information. In the experiments, FCM, FCM with mean spatial information (FCM_S) and fast generalized fuzzy c-means algorithms (FGFCM) are chosen as the comparison methods. Experimental results on two Berkeley images corrupted by noise show the effectiveness of the proposed method.

2 Fuzzy C-Mean Clustering and Its Extended Version

Let $X=\{x_1, x_2, \dots, x_n\}$ denote an image with n pixels, where x_i represents the gray value of the i th pixel. The objective function of standard FCM algorithm is

$$J_m = \sum_{k=1}^K \sum_{i=1}^n u_{ki}^m \|x_i - v_k\|^2 \quad (1)$$

where $v_k (1 \leq k \leq K)$ denotes the center of the k th cluster. $u_{ki} (1 \leq k \leq K, 1 \leq i \leq n)$ represents the membership degree of the i th pixel belonging to the k th cluster and must satisfy $\sum_{k=1}^K u_{ki} = 1$, $u_{ki} \in [0, 1]$, $0 \leq \sum_{i=1}^n u_{ki} \leq n$. In Eq. (1), $\|\cdot\|$ denotes the Euclidean norm and the parameter $m (m > 1)$ is a weighting exponent on each fuzzy membership.

In order to overcome the sensitivity of FCM to noise or artifacts in the image, several modified FCM algorithms utilized the spatial information derived from the image [2-5]. These algorithms first define a spatial constraint term and then introduce it into the objective function of FCM. The modified objective function is

$$J_s = \sum_{k=1}^K \sum_{i=1}^n u_{ki}^m \left(\|x_i - v_k\|^2 + \beta \|\bar{x}_i - v_k\|^2 \right) \quad (2)$$

where \bar{x}_i is the spatial information of the i th pixel. The second term in Eq. (2) is the spatial constraint term and the parameter β controls the penalty effect of the spatial constraint term. The update equations of membership function and cluster center are:

$$u_{ki} = \frac{1}{\sum_{l=1}^K \left(\|x_i - v_l\|^2 + \beta \|\bar{x}_i - v_l\|^2 \right)^{1/(m-1)}} \quad (3)$$

$$v_k = \frac{\sum_{i=1}^n u_{ki}^m (x_i + \beta \bar{x}_i)}{(1 + \beta) \sum_{i=1}^n u_{ki}^m} \quad (4)$$

3 Multiobjective Fuzzy Clustering Algorithm Based on Robust Local Spatial Information

3.1 Chromosome Representation

In general, the gray level value is used as the feature of the pixel in image segmentation. In the multiobjective fuzzy clustering algorithm based on robust local spatial information, the chromosomes are made up of real numbers which represent the cluster centers. The gray level value of image pixel is expressed within a given range between a minimum g_{min} and a maximum g_{max} , therefore each chromosome in the population is randomly initialized in the range between g_{min} and g_{max} . If a particular chromosome encodes the centers of K clusters, its length is taken to be K . For example, the chromosome $<3 58 210>$ encodes three cluster centers: 3, 58 and 210.

3.2 Fitness Function Computation

In the multiobjective fuzzy clustering algorithm based on robust local spatial information, two fitness functions J_s and FS are optimized simultaneously. J_s is the fuzzy compactness with the spatial information. In this paper, the robust local spatial information of the i th pixel [5] is adopted as the spatial information \bar{x}_i in J_s . This spatial information is defined as:

$$\bar{x}_i = \frac{\sum_{p \in S_i} E_{ip} x_p}{\sum_{p \in S_i} E_{ip}} \quad (5)$$

where S_i represents a neighbor window around the i th pixel and E_{ip} denotes a local similarity measure which combines both the spatial coordinates and the gray level values within the neighbor window around the i th pixel. E_{ip} is the weight parameter:

$$E_{ip} = \begin{cases} \exp\left(-\frac{\max(|a_i - a_p|, |b_i - b_p|)}{\lambda_s} - \frac{\|x_i - x_p\|^2}{\lambda_g \sigma_i^2}\right) & i \neq p \\ 0, & i = p \end{cases} \quad (6)$$

where (a_i, b_i) denote the spatial coordinate of the i th pixel. σ_i is defined as

$$\sigma_i = \sqrt{\sum_{p \in S_i} \|x_i - x_p\|^2 / S_R}, \text{ where } S_R \text{ denotes the cardinality of } S_i. \text{ Moreover, } \lambda_s \text{ and } \lambda_g$$

are space and gray scale factors respectively.

FS is the fuzzy separation which can be defined as:

$$FS = \sum_{p=1}^K \sum_{q=1, p \neq q}^K \alpha_{pq}^m \|v_q - v_p\|^2 \quad (7)$$

For FS , if the cluster center v_p is assumed to be the center of the fuzzy set $\{v_q | 1 \leq q \leq K, q \neq p\}$, the membership degree μ_{pq} in Eq. (7) of v_q to v_p is defined as follows:

$$\alpha_{pq} = \frac{1}{\sum_{l=1, l \neq q}^K \left(\frac{\|v_q - v_p\|^2}{\|v_q - v_l\|^2} \right)^{1/(m-1)}}, p \neq q \quad (8)$$

3.3 Selection, Crossover and Mutation

As we already know from the evolutionary algorithms, chromosomes are selected from the population to be parents to crossover and mutation. In the proposed method, crowded binary tournament selection method [12] is adopted to generate the mating pool of chromosomes and the new chromosomes in mating pool are called as parents. In the proposed method, conventional binary tournament method [12] based on the rank and crowded comparison technique is used in the crowded binary tournament selection method.

Crossover and mutation are two basic operators. The double-point crossover operator is utilized in this method. In double-point crossover operator, two crossover

positions are selected uniformly at random and the variables exchanged between the individuals between these cluster centers. Then two new offspring are produced.

After the crossover operator, each center encoded in a chromosome should be mutated with the mutation probability p_m . If the k th cluster center v_k need to be mutated, its value will become $v_k \pm 10\xi$, where ξ is a random number in the range [0, 1] with uniform distribution. Moreover, the '+' or '-' sign occurs with equal probability.

3.4 Elitism and Selection of the Best Solution

As pointed before, NSGA-II is adopted as the underlying multiobjective framework in MFC_RLS to develop the clustering solutions. As is known, the elitism operation is the most characteristic part in NSGA-II. Through elitism operation, the non-dominated solutions among the parent and child populations are propagated to the next generation. Therefore, the good solutions found so far are retained.

In the final generation of MFC_RLS, a set of non-dominated solutions are obtained and all the solutions in final generation are equally important, but sometimes the user may want only one solution. Index I is a recently proposed cluster validity index [13] to produce the final solution. In order to improve the robustness of index I to noise, we introduce the robust local spatial information derived from the image into the index I and propose a clustering index with the robust local spatial information (I_s). Here, I_s is adopted to select a single solution from the non-dominated solutions in the final generation. I_s is defined as follows

$$I_s = \frac{D}{E_s} \quad (9)$$

where $D = \max_{p,q=1}^K \|v_p - v_q\|$ measures the maximum separation between two clusters over all possible pairs of clusters. E_s is defined as

$$E_s = \sum_{k=1}^K \sum_{i=1}^n u_{ki} [\|x_i - v_k\| + \|\bar{x}_i - v_k\|] \quad (10)$$

These two factors are found to compete with and balance each other critically. Larger value of index I_s implies the better solution obtained by the method.

4 Experimental Results and Analysis

To verify the effectiveness of MFC_RLS, FCM, FCM_S [2] and FGFCM [5] are adopted as the comparative methods. In order to evaluate the algorithm performance, the segmentation accuracy (SA) [15] and adjusted rand index (ARI) [16] are used. In addition, two real images chosen from the Berkeley Segmentation Dataset [24] (Shown in Figs. 1(a) and 2(a)) are used in this experiment. For all the methods, the number of clusters is set by manual. The fuzziness index m is set to 2. For FCM,

FCM_S and FGFCM, the maximal iteration num T and the stopping threshold ε are set to 300 and 10^{-5} , respectively. For FCM_S, FGFCM and MFC_RLS, the size of the neighbor windows is 3×3 . The parameters λ_s and λ_g in FGFCM and MFC_RLS are set to 3 and 6. Moreover, the spatial constrained parameter β in FCM_S and MFC_RLS is set to 6. The evolutionary parameters of MFC_RLS used in our experimental study are set as follows: the maximum number of generations is 50, the population size is 20, the crossover probability is 0.9, and the mutation probability is 0.1.

Table 1. Comparison of these four methods on Berkeley images corrupted by Gaussian noise

Image	#3096	#238011
K	2	3
Gaussian Noise Level	(0, 0.008)	(0, 0.01)
FCM	SA	0.6736
	ARI	0.0802
FCM_S	SA	0.9202
	ARI	0.4984
FGFCM	SA	0.9466
	ARI	0.6097
MFC_RLS	SA	0.9718
	ARI	0.7499

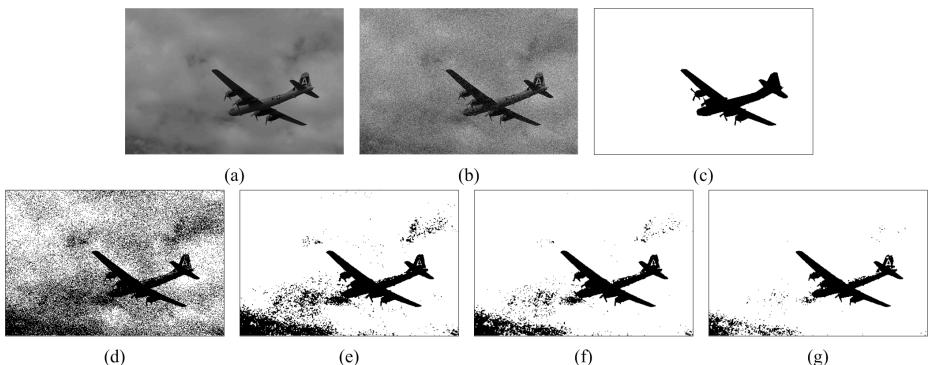


Fig. 1. Segmentation results on the image #3096 corrupted by the Gaussian noise (0, 0.008): (a) original image; (b) noisy image; (c) Benchmark image; (d) FCM; (e) FCM_S; (f) FGFCM; (g) MFC_RLS.

In order to investigate the algorithm performance, Gaussian white noise is artificially added to these two images and the corresponding noisy images are shown in Figs. 1(b) and 2(b). For each image, the benchmark image (Shown in Figs. 1(c) and 2(c)) is segmented by manual. Table 1 presents the SA and ARI values of different algorithms on these two noisy images and the results reveal that MFC_RLS obtains the highest SA and ARI values among all the methods. In order to more obviously

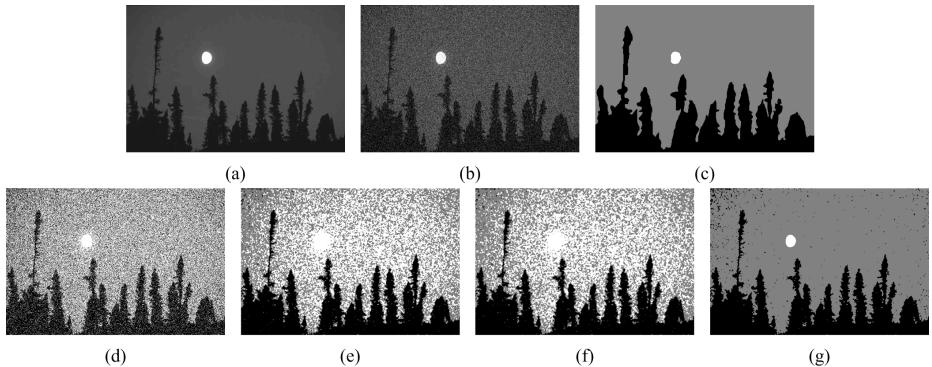


Fig. 2. Segmentation results on the image #238011 corrupted by the Gaussian noise (0, 0.01): (a) original image; (b) noisy image; (c) Benchmark image; (d) FCM; (e) FCM_S; (f) FGFCM; (g) MFC_RLS.

illustrate the visual results, the segmentation results on these two images are shown in Figs. 1-2 (d)-(g). For the image #3096, the visual results of FCM_S, FGFCM and MFC_RLS are almost the same. However the noise in the result of MFC_RLS is the least. Hence, it is evident that the proposed method outperforms all the comparison methods in the quantitative evaluation and visual effect. For the image #238011, due to considering the fuzzy separation, it is obvious that the proposed MFC_RLS method produces higher SA and ARI values than the other algorithms. Moreover, MFC_RLS also obtains the best visual result on this image.

5 Conclusions

In this article, a multiobjective fuzzy clustering algorithm based on robust local spatial information (MFC_RLS) for image segmentation is proposed. In the proposed method, the robust local spatial information derived from the image is introduced into the fitness function and cluster validity index. The improved fitness function and cluster validity index overcome the influences of image noise to the clustering performance. The experimental results show that MFC_RLS outperforms FCM, FCM_S and FGFCM.

In future works, in order to overcome the influence of higher level noise to algorithm performance, some more effective spatial information can be incorporated into the proposed method. In addition, the number of clusters is set by manual in this paper, how to automatically evolve is also our future research.

References

1. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley, Massachusetts (1992)
2. Chen, S.C., Zhang, D.Q.: Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. IEEE Transaction on System, Man, and Cybernetics, Part B: Cybernetics 34(4), 1907–1916 (2004)

3. Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A., Moriarty, T.: A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Transaction on Medical Imaging* 21(3), 193–199 (2002)
4. Szilagyi, L., Benyo, Z., Szilagyii, S., Adam, H.S.: MR brain image segmentation using an enhanced fuzzy C-means algorithm. In: Proceedings of 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 724–726. IEEE Press, Cancun (2003)
5. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition* 40(3), 825–838 (2007)
6. Zhao, F., Jiao, L.C., Liu, H.Q.: Fuzzy c-means clustering with non local spatial information for noisy image segmentation. *Frontiers of Computer Science in China* 5, 45–56 (2011)
7. Handl, J., Knowles, J.: An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation* 11(1), 56–76 (2006)
8. Mukhopadhyay, A., Maulik, U.: A multiobjective approach to MR brain image segmentation. *Applied Soft Computing* 11(1), 872–880 (2011)
9. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York (1981)
10. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 841–847 (1991)
11. Pakhira, M., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. *Pattern Recognition* 37, 487–501 (2004)
12. Deb, K., Agrawal, S., Pratab, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) *PPSN 2000. LNCS*, vol. 1917, pp. 849–858. Springer, Heidelberg (2000)
13. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(12), 1650–1654 (2002)
14. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceeding of the 8th International Conference of Computer Vision, vol. 2, pp. 416–423 (2001)
15. Wu, M., Scholkopf, B.: A local learning approach for clustering. In: *Advances in Neural Information Processing Systems*, pp. 1529–1536. MIT Press, USA (2007)
16. Yeung, K.Y., Ruzzo, W.L.: An empirical study on principal component analysis for clustering gene expression data. *Bioinformatics* 17(9), 763–774 (2001)

Image Super-Resolution Based on Data-Driven Gaussian Process Regression

Yan-Yun Qu, Meng-Jie Liao, Yan-Wen Zhou,
Tian-Zhu Fang, Li Lin, and Hai-Ying Zhang

Computer Science Department, Xiamen University, 361005, China
quyanyun@gmail.com

Abstract. In this paper, we aim at producing the super-resolution image from a single low-resolution image based on Gaussian Process regression. Gaussian Processes provide a framework for deriving regression techniques with explicit uncertainty models. Super resolution can be transformed into a regression problem. We show how Gaussian Processes with covariance functions can be used for image super-resolution. Furthermore, considering that the training data have greatly effect on the super-resolution performance and the unsuitable training data would result in unexpected details, we adopt a data-driven scheme to learn a regression map for each query patch. There are two advantages of our approach: 1) we establish a map between the low-resolution space and the high-resolution space independent of a specified regression function; 2) the data-driven learning scheme improves the super-resolution performance. We estimate our approach on the popular testing images which are used in other super-resolution literatures, and the results demonstrate that our approach is efficient and it manifests a high-quality performance compared with several popular super-resolution methods.

Keywords: Super Resolution, Gaussian Process Regression, Covariance Matrix, Data-driven.

1 Introduction

Image super-resolution (SR) has been a hot topic in the field of computer vision, and it's widely used in many practical applications, such as satellite imaging and medical image formation where the analysis or diagnosis from low-resolution (LR) images can be very tough. The goal of image SR is to generate a high-resolution (HR) image from one or a set of low-resolution (LR) input images. The SR problem is ill-posed because a low-resolution image can be generated by many different high-resolution images under different transformations. There are many literatures to discuss image super-resolution, which can be divided into three classes: the interpolation based methods, the reconstruction based methods and the learning-based methods. Interpolation-based methods [1] [4] [5] are simple and fast, but the quality of the super-resolution image is very limited, because they cannot recover the high frequency details. Reconstruction based

methods [2] [6–10] apply various smoothness priors and impose the constraint that makes the HR image reproduce the original LR image when properly down-sampled. The limit of these methods is that they require that the smoothed and downsampled version of an HR image should be similar to the original LR image. Their performance relies on the prior information and the compatibility with the given image. Moreover, their performance degrades rapidly with the increase of the magnification factor, or with the decrease of the size of the input image.

Alternatively, the learning based methods [11–14] are promising, where detailed textures are hallucinated by searching through a training set which contains pairs of LR and HR patches. Freeman [11] proposed an example-based learning approach in which the prediction from the low-resolution image to the high-resolution image was learned via a Markov Random Filed (MRF). Yang [15] transformed the super-resolution problem into a sparse representation problem. Instead of dealing with image patches, Sun [3] proposed to use the primal sketch priors to enhance blurred edges.

In learning based SR methods, we are concerned about the regression based SR methods. Ni [16] used support vector regression (SVR) to solve the super-resolution problem in the frequency domain. Kim [19] estimated the high frequency details based on the kernel ridge regression. In these methods, the regression functions were specified. And the super-resolution performance depends on whether a specified regression model is suitable. In this paper, we want to adopt a finer method than those specified regression models, in which the regression function is represented by the data. Thus, we utilize Gaussian Process regression to solve the super-resolution problem, because Gaussian Processes provide a framework for deriving regression techniques with explicit uncertainty models. He [17] showed the feasibility of Gaussian Process regression to solve the SR problem, but in his method, a HR image is produced from a single image without the external training dataset. Our method will use Gaussian Process regression to learn the relationship between the LR space and the HR space from the training data. Furthermore, we investigate the training data selection. In the learning based method, the training data should be selected carefully, otherwise the unexpected details are introduced. Thus, we implement a data-driven scheme to learn a regression map for each query image patch, that is, for a query image patch, we collect its nearest neighbors as the training data.

Our contributions are as follow: 1) we use Gaussian Process regression to solve the super-resolution problem, which avoid specifying a special regression model; 2) we implement a data-driven regression scheme, in which each query image patch has its special training data and its special regression model.

The remainder of this paper is organized as follows: we give a brief overview of Gaussian Process regression and describe our algorithm of our approach in Section 2. Section 3 presents the experimental results, and conclusions are given in Section 4.



Fig. 1. Some images in the training collection

2 SR Framework Based on Gaussian Process Regression

The overall processing pipeline of our approach is as follows. Given an LR image, we firstly interpolate it into the desired scale , and then we use the learned fitting map to produce an HR image which recovers the missing high frequency details. Our approach adopts the framework of Freeman’s work [11]. But the learning relationship is different between our approach and Freeman’s work, our approach learns the relationship between the HR images and the corresponding interpolated images of their LR image named the blurred HR images while Freeman’s work learns the relationship between the high frequency detail images and the LR images.

2.1 Training Set Generation

We download 100 natural images from the Internet to make the data set collection. The downloaded image collection consists of four types of images: animals, plants, buildings and people. Some examples are shown in Fig.1. Then we perform a down-sampling process on each image, and the obtained low-resolution image and its high-resolution image form a training pair. Because we want to learn the map between the space of blurred HR images and the space of HR images, we initially apply the interpolation method, such as the nearest interpolation method in this paper, to the LR image, by which we obtain the blurred HR images. Then we randomly sample the patches in an HR image and the corresponding patches in the blurred HR image with the same size (N pixels, or $\sqrt{N} \times \sqrt{N}$) in the image collection and obtain the training pairs that include the HR patches and the corresponding blurred HR patches. The input of the regression map is the patches from the blurred HR images that form a set $X = \{x_1, x_2, \dots, x_n\}$, and the output of the regression map is the patches from the HR images that form a set $Y = \{y_1, y_2, \dots, y_n\}$. In this paper, the magnification factor is 4, that is, the magnification factor for the width and the height is 2 respectively. We may implement another downsampling factor on all images in the collection to generate other scales of degradation corresponding LR images.

2.2 Gaussian Process Regression

As we know, a regression problem usually require a specified function $f(x)$, for example, quadratic, cubic and nonpolynomial. If the specified function is suitable to the data, we will obtain a high quality solution. However, it is not a trivial task to make it clear if the function $f(x)$ is suitable to the data. A Gaussian Process can represent $f(x)$ obliquely, but rigorously, by letting the data decide the function. Gaussian Process extends the multivariate Gaussian distribution to infinite dimensionality, and it defines a distribution on the function f . Joint probability distribution of any finite subset of random variables x , $\{x_i \in X, i = 1, 2, \dots, n\}$ and its corresponding process state y , $\{y_i \in Y, i = 1, 2, \dots, n\}$ obeys n -dimensional Gaussian distribution. Gaussian Process is parameterized by its mean function $m(x)$ and covariance function $k(x_i, x_j)$ which is the element of the matrix $K(X, X)$ such that $f(x) \sim \mathcal{GP}(m(x), k(x_i, x_j))$. Let y denote one observation with Gaussian noise ε and we have the Gaussian Process regression model $y = f(x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$.

The covariance function can be decided by the task and it should produce a non-negative definite covariance matrix. Any finite dimension distribution for Gaussian Process is still Gaussian distribution. The joint distribution of the training outputs y and the test outputs y' with a zero mean function is

$$\begin{bmatrix} y \\ y' \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K(X, X) + \sigma_n^2 & K(X, x') \\ K(x', X) & K(x', x') \end{bmatrix}). \quad (1)$$

where X and x' are design matrices for training data and test data respectively. Conditioning y' on the observation y , we compute the predictive distribution $y' | X, y, x'$ as $y' | X, y, x' \sim \mathcal{N}(\bar{y}', Var(y'))$, where $\bar{y}' = K(x', X)[K(X, X) + \sigma_n^2 I]^{-1}y$, and $Var(y') = K(x', x') - K(x', X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, x')$.

2.3 Covariance Function for Gaussian Process Regression

Considering Gaussian process is parameterized by its mean function $m(x)$ and covariance function $k(x_i, x_j)$, the choice of covariance function will have a great effect on the prediction accuracy. We model the image as a locally stationary Gaussian Process and choose the squared exponential covariance function $k(x_i, x_j) = \sigma_f^2 \exp(-\frac{1}{2} \frac{\|x_i - x_j\|^2}{l^2})$. where σ_f^2 represents the signal variance and l^2 defines the characteristic length scale.

Because the covariance function is a symmetric positive definite function, the covariance function can be seen as the kernel function. For nonlinear problems in the input space, we use the kernel function of the input space to implicitly map to a high dimensional feature space.

2.4 Optimizing Hyper-Parameters and Covariance Matrix

In this paper, there are three hyper-parameters σ_n , l and σ_f . The value of the hyper-parameters plays the central role on Gaussian process. For each query LR

Algorithm 1. SR based on Data-driven GPR

```

1: Initializing  $InputLR$  being an input LR image ;
2: Nearest interpolation:  $InterpLR \leftarrow InputLR \uparrow^m$ ;
3: Partition  $InputLR$  into  $n$  overlapped patches  $P_1, \dots, P_n$ ;
4: for each  $P_L = P_1, \dots, P_n$  do
5:   Search  $n$  nearest neighbors of  $P_L$  to construct the training data
    $\{\langle P_L^1, P_H^1 \rangle, \dots, \langle P_L^n, P_H^n \rangle\}$  ;
6:   Put the  $n$  pairs of patches into  $trainLset$  and  $trainHset$  as training vectors;
7:   Train a GPR covariance matrix  $K(X, X)$  using  $trainLset$ ;
8:   Optimize hyper-parameters  $\{\sigma_f, l, \sigma_n\}$  and covariance matrix;
9:   Train a GPR model  $M$  using  $\{\sigma_f, l, \sigma_n, K(X, X)\}$ ;
10:   $P_H \leftarrow M(P_L)$ ;
11: end for
12: return  $OutputHR$  constructed from  $P_H$ ;

```

image patch, we should optimize these three hyper-parameters. Here we explore how to set them through optimization.

Let θ be the vector composed of these hyper-parameters in covariance function. According to Bayes theorem, we can get $p(\theta|y, X) = \frac{p(y|X, \theta)p(\theta)}{p(y|X)}$, and $p(\theta|y, X) \propto p(y|X, \theta)p(\theta)$. Under the assumption that θ approximately obey the uniform distribution, it is equally likely no matter what values of hyper-parameter to be, so

$$\arg \max(p(\theta|y, X)) \approx \arg \max(p(y|X, \theta)), \hat{\theta} = \arg \max(p(y|X, \theta)),$$

where $p(\theta|y, X)$ is the marginal likelihood and the logarithmic form is

$$\log p(y|X, \theta) = -\frac{1}{2}y^T K^{-1}y - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi.$$

In our experiments, the initial value of σ_f is the standard deviation of current input LR image patch x' . And the initial value of l is equal to 0.223, and σ_n is equal to 0.01. Then we use the gradient descent method to get the maximum value of $\log p(y|X, \theta)$. When $\log p(y|X, \theta)$ obtains the maximum value, $\hat{\theta}$ is the estimated value of θ .

2.5 Gaussian Process Regression for SR

In our regression-based framework, patches from the LR image are predicted by the Gaussian Process regression. Let the input LR image be regarded as $InputLR$, we firstly interpolate it into the desired scale by traditional interpolation method, such as the nearest interpolation method, and we obtain the blurred HR image($InterpLR$) which loses high-frequency details. In this paper, the relationship is established between the blur HR space and the HR space.

Now we consider the construction of the covariance matrix. In order to achieve the good SR performance, we select the proper training data in a data-driven

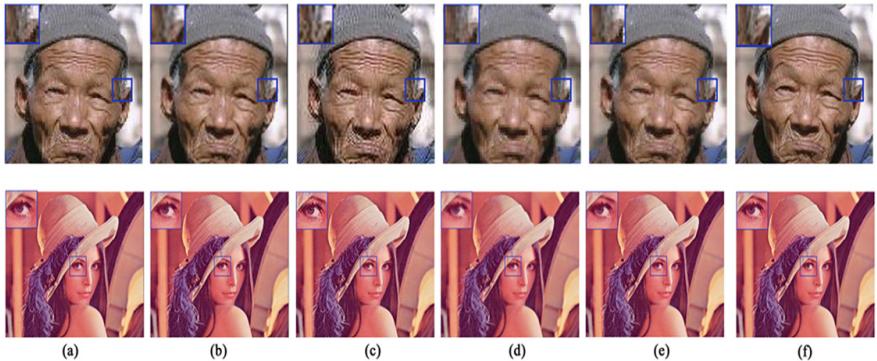


Fig. 2. Comparison results of different SR methods. (a)Nearest neighbor interpolation. (b)Bicubic interpolation. (c)Sparse representation. (d)GPR. (e)DDGPR. (f)Ground truth image.

way. For a blurred HR image, we search its similar patches in the training dataset, and then we compute the value of $k(x_i, x_j)$ as the (i, j) element value of the covariance matrix $K(X, X)$. Then an HR image patch is generated by an optimized covariance function.

As shown in Algorithm 1, the algorithm carries out the training and reconstruction for each query image patches (*InputLR*). And different image patches corresponds to different covariance matrixes. In our approach, we use the Gaussian Process regression to predict each pixel for the HR image. In Equ.1, the (i, j) element of the training covariance matrix $K(X, X)$ is $k(x_i, x_j)$, and the matrix $K(X, x')$ is the transposition of the matrix $K(x', X)$, where $K(x', X)$ is defined as $K(x', X) = [K(x', x_1), K(x', x_2), \dots, K(x', x_n)]$, x' is the blurred HR image of a query LR image patch, and X is the LR training data. The vector y consists of n elements, and each element is the pixel values in the same position of n HR training patches, and y' is the pixel value in the corresponding position of the HR image of the query image.

3 Experimental Results

In this paper, we build the training dataset which contains 20,000 pairs of the LR images and the HR images and the magnification factor is 4. The parameters of l and σ_n are 0.223, 0.01 respectively. For each query image patch, we only search 30 pairs of patches from the training data whose LR patches are the most similar to the current LR image patch.

We compare our method with the nearest neighbor interpolation, the bicubic interpolation and the sparse representation method. Fig.2 shows the SR results of the six methods at the magnification factor 4 for two test images. Our method can achieve the best SR performance than other methods in terms of the visual effect.

In order to estimate the effect of the data-driven scheme, we compare our method with the standard Gaussian Process regression SR which establish only a covariance matrix for a predict pixel value of the HR image patch. We call the method based on data-driven Gaussian Process regression DDGPR, and call the SR method only based on Gaussian Process regression GPR. So in the second experiment, we implement GPR on the testing images. In details, we firstly cluster training data (*contained 20,000 pairs*) into k groups by K-means method. Secondly, we randomly select m samples from every group and use these samples to train a GPR model (*including covariance matrix, σ_n*). We can get a new training dataset which includes k GPR models. When a query LR image patch is input, we first find its group. The query LR image patch will belong to the group whose center has the smallest distance to the query LR image patch among the k groups. We utilize the learned GPR model from this group of training data to produce the HR image patch. So we can reconstruct the HR image at a high computational speed. In our experiment, the parameters of k and m are 200, 50 respectively. The performance comparisons in terms of PSNR are shown in Table 1. For simplicity, Nearest stands for the nearest neighbor interpolation, Bicubic stands for the bicubic interpolation, Sparse stands for the SR via sparse representation method[18]. Our approaches achieve the highest PSNR values. Therefore, our approaches are superior to the other four methods.

Table 1. PSNR Comparison of Different SR Methods

Image	Nearest	Bicubic	Sparse	DDGPR	GPR
Child	20.2532	21.3129	24.3358	25.2050	25.0667
Mantis	24.9631	26.0007	28.6380	30.5287	30.3183
Lena	26.6390	28.7416	28.9703	29.8040	29.6334
Builfing	18.9651	20.6895	19.1884	20.6636	20.3299
Old man	26.3307	25.4827	26.1263	27.4873	27.1649
Sculpture	30.0020	26.2252	28.9550	30.0803	29.8831

4 Conclusions

In this paper we present a novel method for a single image super-resolution based on data-driven Gaussian Process regression. We transform the SR problem to a regression problem, but we do not specify a regression function. Gaussian Process regression make the data represent the regression function. Moreover, we implement a data-driven scheme to learn a relationship for each query image between the LR images and HR images. The experimental results demonstrate that our approach is superior to the bicubic interpolation method, the nearest neighbor interpolation method and the sparse representation method. DDGPR achieves the best visual effect of SR, and achieves the highest PSNR values.

Acknowledgments. This research work was supported by the Fundamental Research Funds for the Central Universities (2010121067), the Natural Science Foundation of Fujian Province of China (2013J01257,2013J01249), 2013 national college students' innovative and entrepreneurial training project and "A" plan of the education department of Fujian Province of China (JA11289).

References

1. Xin, L., Orchard, M.T.: New edge directed interpolation. In: Processing International Conference, pp. 311–314 (2000)
2. Schultz, R.R., Stevenson, R.L.: A Bayesian approach to image expansion for improved definition. *IEEE Trans. on Image Processing* 3, 233–242 (1994)
3. Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: CVPR 2008, Anchorage, AK, United states (2008)
4. Hou, H.S., Andrews, H.: Cubic splines for image interpolation and digital filtering. *IEEE Trans. on Acoustics, Speech and Signal Processing* 26, 508–517 (1978)
5. Tam, W.S., Kok, C.W., Siu, W.C.: Modified edge-directed interpolation for images. *Journal of Electronic Imaging* 19, 013011-013011-20 (2010)
6. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. *IEEE Trans. on PAMI* 24, 1167–1183 (2002)
7. Irani, M., Peleg, S.: Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation* 4, 324–335 (1993)
8. Kwok-Wai, H., Wan-Chi, S.: New motion compensation model via frequency classification for fast video super-resolution. In: Image Processing, pp. 1193–1196 (2009)
9. Morse, B.S., Schwartzwald, D.: Image magnification using level-set reconstruction. In: CVPR 2001, vol. 1, pp. 333–340 (2001)
10. Shan, Q., Li, Z., Jia, J., Tang, C.-K.: Fast image/video upsampling. *ACM Transactions on Graphics*, 153 (2008)
11. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. *IEEE Computer Graphics and Applications* 22, 56–65 (2002)
12. Hong, C., Dit-Yan, Y., Yimin, X.: Super-resolution through neighbor embedding. In: CVPR 2004, vol. 1, pp. 275–282 (2004)
13. Liu, C., Shum, H.Y., Freeman, W.T.: Face hallucination: Theory and practice. *International Journal of Computer Vision* 75, 115–134 (2007)
14. Qiang, W., Xiaoou, T., Shum, H.: Patch based blind image super resolution. In: Computer Vision, vol. 1, pp. 709–716 (2005)
15. Jianchao, Y., Wright, J., Huang, T., Yi, M.: Image super-resolution as sparse representation of raw image patches. In: CVPR 2008, pp. 1–8 (2008)
16. Ni, K.S., Nguyen, T.Q.: Image superresolution using support vector regression. *IEEE Trans. on Image Processing* 16, 1596–1610 (2007)
17. He, H., Wan-Chi, S.: Single image super-resolution using Gaussian process regression. In: CVPR 2011, pp. 449–456 (2011)
18. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE Trans. on Image Processing* 19, 2861–2873 (2010)
19. Kwang In, K., Younghhee, K.: Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Trans. on PAMI* 32(6), 1127–1133 (2010)

Face Recognition Based on Non-Subsampled Contourlet Transform and Multi-order Fusion Binary Patterns

Yao Deng, Weifeng Li, Zhenhua Guo, and Youbin Chen

Graduate School at Shenzhen, Tsinghua University, China

thuyao2011@gmail.com,

{li.weifeng,zhenhua.guo,chenyb}@sz.tsinghua.edu.cn

Abstract. In this paper, we propose a novel face representation approach based on Non-subsampled Contourlet Transform (NSCT) and Multi-order Fusion Binary Patterns (MFBP). NSCT, which is a newly developed multi-resolution analysis tool in image denoising and enhancement, can be used to effectively capture image features of both geometrical structure and directional texture information. Due to the ability of extracting multi-order derivatives of texture patterns, the MFBP is applied on the NSCT coefficient images to achieve enhanced face representation. Furthermore, Block-based Fisher Linear Discriminant (BFLD) feature selection and weight scheme based on Fisher Separation Criteria (FSC) are chosen to further improve discriminative power of the proposed face representation. The experiments on public FERET database demonstrate that our approach outperforms many of the state-of-the-art methods.

Keywords: Non-subsampled contourlet transform (NSCT), multi-order fusion binary patterns (MFBP), block-based Fisher linear discriminant (BFLD), face recognition.

1 Introduction

Face recognition, as one of the most important biometric technologies, has been extensively studied over the last few decades. Due to its advantages of being natural, non-intrusive, and convenient to acquire samples, automatic face recognition has the broad application prospect in a variety of areas such as biometric authentication, video surveillance, law enforcement, and so on. Numerous approaches have been developed for face recognition and many of them achieved satisfying results under the controlled circumstances. Nevertheless, the automatic face recognition under uncontrolled conditions remains very challenging for the reason that large intra-personal variations and small inter-personal variations are caused by uncontrolled illumination, expression, pose, aging and so on. Therefore, seeking more effective face representation approaches becomes the research focus.

Among the well-known face representation methods, local binary pattern (LBP) has proved to be an efficient texture descriptor, which is firstly applied to face descriptor by Ahonen et al. [1]. Improvements of LBP for face recognition include local derivative pattern (LDP) [2], local XOR pattern (LXP) [3], and so on. However,

according to the latest research [4][5], applying local pattern operators on decomposed images of frequency analysis or wavelet analysis achieves better performance than applying them on gray-scale images. Gabor wavelets are multi-resolution and multi-directional analysis tools, so they can easily describe more discriminative information of images. As we know, Gabor wavelets have been one of most successful features in face recognition domain. And high recognition accuracy on public face databases attests to effectiveness of face representation based on Gabor and LBP (or methods close to them).

According to recent study [6], Gabor may not be the best feature in terms of face representation. As a newly developed multi-resolution analysis tool, Non-subsampled Contourlet Transform (NSCT) [7] can also describe images in multiple scales and directions. Due to its anisotropy, shift-invariance, and intrinsic geometrical description, NSCT can achieve better performance than Gabor wavelets. In this paper, we dig deeper to discuss the potential of NSCT in face representation. To fully utilize information, we propose a set of fusion patterns called multi-order fusion binary patterns (MFBP), which fuse multi-order derivatives of texture information. The proposed method enhances the expressive power of original LBP and achieves better performance. More importantly, MFBP have simpler form than complex LBP variants such as LDP [2]. To further probe into the face representation ability of NSCT, a new face representation approach based on NSCT and MFBP is proposed. The experiments on public FERET database demonstrate that our approach outperforms many of the state-of-the-art methods, including Gabor wavelet based methods.

The paper is organized as follows. Section 2 gives the overview of NSCT and LBP. Section 3 introduces the proposed MFBP descriptor, and presents the proposed framework utilizing our texture descriptor on NSCT decomposed images and block-based Fisher linear discriminant (BFLD). Then Section 4 reports experimental results to illustrate the effectiveness our method. Section 5 concludes the whole paper.

2 Brief Review of NSCT and LBP

2.1 Non-Subsampled Contourlet Transform (NSCT)

Contourlet transform was firstly introduced by Do and Vetterli in literature [8]. It is designed to be multi-scale and multi-directional with Laplacian pyramid and directional filter banks, which aim to represent the 2D geometry of digital images. However, contourlet transform is not robust to noises, and more importantly, it is not shift-invariant. Based on a non-subsampled pyramid structure and non-subsampled directional banks, Cunha et al. proposed non-subsampled contourlet transform (NSCT) [7], which can overcome the aforementioned shortcoming of contourlet transform. An example of decomposing a facial image in two scales and 4, 8 directions respectively is illustrated in Fig. 1 (the low-pass sub-band result doesn't count as one scale). As shown in Fig. 1 also, the frequency plane in the sub-bands is spilt by the directional filter banks. More details can be found in [7]. As we can see in the decomposed images, different directional edges of facial components are well preserved in the first scale, while weaker edges of face images are captured in the second scale.

These micro-patterns of facial images contain discriminant information. Moreover, they are illumination-invariant. Because of the rich basis functions oriented at various directions and multiple scales, NSCT can effectively extract smooth contours which are the dominant features in facial images. NSCT coefficient images can be used to extract robust binary patterns of texture, which will be introduced in the following parts of the paper.

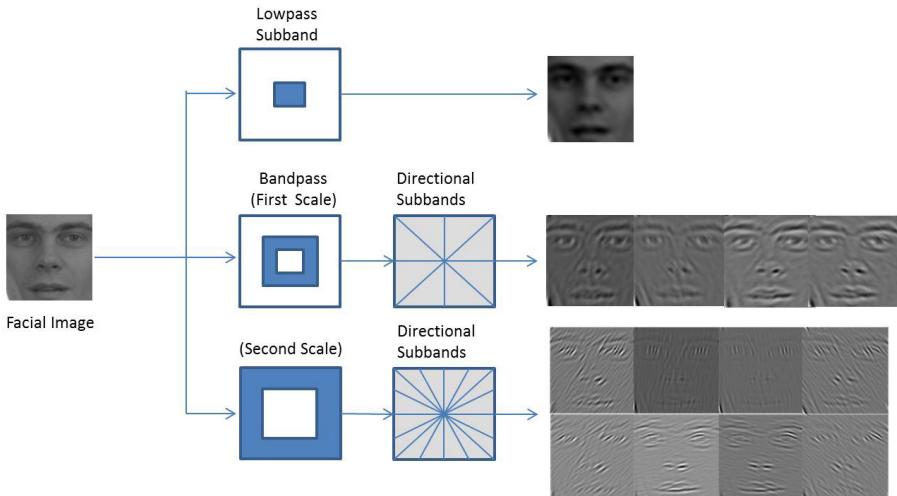
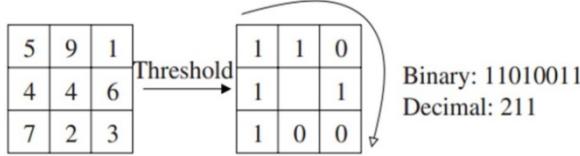


Fig. 1. An example of Non-subsampled Contourlet Transform for facial images in two scales and 4, 8 directions in the scales respectively

2.2 Local Binary Patterns (LBP)

In face recognition domain, LBP has been proved to be an efficient texture descriptor. It is designed to represent local patterns by thresholding k -by- k neighborhood of each pixel with the center pixel value, and concatenating these results to form a binary number. Extracting local binary patterns can actually be seen as the following steps: first, compute 8 directional derivative pictures on the original image; second, binarize each pixel of derivative pictures to “1” or “0” according to its value; finally, get the LBP image by adding all binarized derivative images using different power of 2 as weight values. An example of LBP calculation is shown in Fig. 2 [1]. The formula for LBP example of 3-by-3 neighborhood is shown in (1), where I_n represents the intensity of pixels in neighborhood and I_c means the intensity of the central pixel.

Ojala et al. [9] put forward that 8-neighborhood LBP has a total of 256 patterns, however, “uniform patterns” account for a bit less than 90% of all patterns. In a binary string, if the shifts between 0 and 1 are less than or equal to 2, it is called “uniform pattern”, such as: 00000000, 00011110, and 10000011. These uniform patterns contain most of rich discriminative information.

**Fig. 2.** An example of basic LBP operator

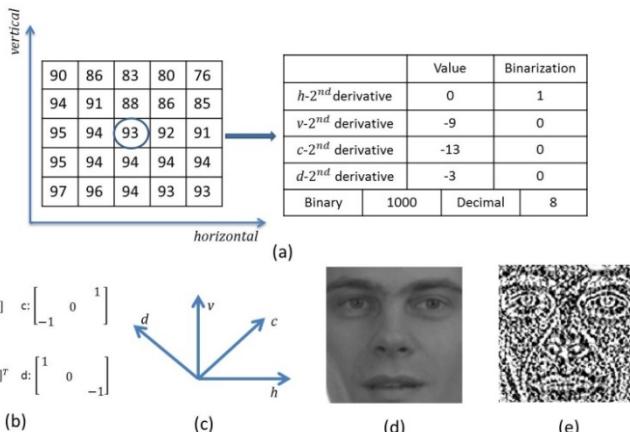
$$LBP(x_c, y_c) = \sum_{n=0}^7 2^n s(I_n - I_c). \quad (1)$$

Following the work in [1], we use the notation $LBP_{P,R}^{u2}$ in this paper. The subscripts represent that P sampling points are selected on a circle of radius of R. The superscript means using only uniform patterns and labeling all remaining patterns with a single label.

3 Proposed Face Recognition Approach

3.1 Multi-order Fusion Binary Patterns (MF BP)

Take gray-scale images as an example, derivatives of image texture have explicit physic meaning: first order derivatives in some direction indicate the change of gray-scale along this direction; and second order derivatives reveal intensity of that change. As traditional LBP only capture 1st-order texture derivative features, in this paper, to explore image information more deeply, we proposed to combine high-order texture information with widely used local binary features.

**Fig. 3.** Illustration of 2nd-order binary patterns (LBP^2). (a) An example of computing 2nd-order derivatives of four directions. (b) Selected operators for four directions. (c) Four directions. (d) An example of original face image. (e) Visualization of LBP^2 .

Inspired by Gabor Surface Feature in [10], we use simple symmetric derivative operator [-1,0,1]. To extract 2nd-order derivative information, four directions are chosen: horizontal, vertical, diagonal, and counter-diagonal. The operator variants in different directions are illustrated in Fig. 3(b). The workflow of the proposed MFBP can be described as follows: first, if the original image is denoted as I , for all pixels in the image, 2nd-order derivatives in some direction can be computed by applying the derivative operator twice, the results of all pixels are seen as the 2nd-order directional derivative picture; second, using the threshold “0” for binarization, the 2nd-order binary patterns (LBP^2) can be formulated in (2), where I''_x means the binarized results of 2nd-order directional derivative pictures, the superscript denotes the order of derivatives and the subscript “ x ” denotes the directions, which are displayed in Fig. 3(c); after that, the histogram of our proposed MFBP is formulated in (3), where “ $H(\cdot)$ ” means calculating the histogram and “ \oplus ” refers to concatenation of histograms. Finally, the proposed MFBP only have $59+16=75$ dimensions in a histogram. For clarity, an example of extracting one pixel’s 2nd-order binary pattern value is shown in Fig. 3(a).

$$LBP^2(I) = 2^3 I''_h + 2^2 I''_v + 2^1 I''_c + 2^0 I''_d, \quad (2)$$

$$MFBP_histogram(I) = H(LBP_{(8,2)}^{u2}(I)) \oplus H(LBP^2(I)), \quad (3)$$

3.2 MFBP on Decomposed Images of NSCT and Weighted Block-Based FLD

Recent studies [4][5] have shown that local patterns of frequency domain or wavelet domain are much more robust to various variations. This strategy has two advantages: firstly, frequency domain or wavelet domain which utilizing multi-resolution analysis can capture much richer image features robustly; secondly, converting facial images to spatial histograms of local patterns can not only reduce dimensionality, but also overcome side-effect of face misalignment. Therefore, we apply the proposed MFBP with a subsequent spatial structure histogram model on decomposed images of NSCT.

Firstly, we decompose a facial image into overall 21 pictures of three scale levels [6]; Secondly, each decomposed picture is divided equally into 4*4 non-overlapped blocks and then each block is further partitioned into 2*2 sub-regions, a MFBP histogram is computed from every sub-region; After that, we concatenate all MFBP histograms within each block (not sub-region), all decomposed images are then concatenated into 16 concatenating histograms.

A lot of works have demonstrated that different parts of facial image have different ability of identification. So, different blocks should be assigned different weights. Here, we propose to adopt weighting scheme based on Fisher Separation Criteria (FSC) [3]. For each block, the weighting criteria can output a score as its weight according to its intra-class similarities and inter-class differences.

Following the aforementioned rule, we divide all decomposed images into 4*4 blocks for Block-based Fisher Linear Discriminant (BFLD) [3][4]. Here, for the general case, we use variable *numblock* (means the number of blocks) to replace 16. Each block has a concatenated histogram of MFBP, so we have total histogram sequence:

$$V = (H_1, H_2, H_3, \dots, H_{numblock}), \quad (4)$$

Then we use BFLD projection matrices to transform H_i to low-dimensional X_i , W_i^{BFLD} indicates the i -th block's projection matrix,

$$X_i = (W_i^{BFLD})^T H_i, (i = 1, 2, 3, \dots, numblock), \quad (5)$$

The matching score for two facial images I and I' can be formulated as:

$$score(I, I') = \sum_{i=1}^{numblock} fscw_i * S(X_i, X'_i), \quad (6)$$

where $fscw_i$ denotes the i -th block weight by FSC, and $S(x, y)$ means the distance metric for vectors, here we use cosine distance.

4 Experiments

4.1 Experiment Setting

Experiments are conducted on FERET public face database to demonstrate the effectiveness of our proposed approach. All face images are properly aligned according to localization coordinates of both eyes. And no further preprocessing is applied.

We test our method on the public database through standard FERET protocol [11]. The gallery set fa consists of 1196 images of 1196 subjects. There are four probe subsets for frontal face recognition: fb subset (subset of expression, 1195 images), fc subset (subset of illumination, 194 images), dupI subset (subset of time I, 722 images are taken later in time), dupII subset (subset of time II, images are taken at least one year after the corresponding gallery images, it is actually a subset of dup I).

4.2 NSCT+MFBP vs. Gabor or NSCT+LBP

In this section, we want to test the effectiveness of our proposed framework. So we conduct experiments of four methods: 1) LBP of Gabor magnitude images + BFLD, 2) MFBP of Gabor magnitude images + BFLD, 3) LBP of NSCT images + BFLD, 4) MFBP of NSCT images + BFLD. The partition scheme of facial images is the same as mentioned before. The standard training set, which consists of 1002 images of 429 subjects, is used for training FLD projection matrix. LBP refers commonly used $LBP_{8,2}^u$ in [1], which has 59 bins for each histogram calculation. The results of all test sets are illustrated in Table 1.

From Table 1, we can see that the performance of proposed NSCT+MFBP framework is better than other combinations. It shows that NSCT is more appropriate for face recognition than Gabor wavelet. Table 1 also shows that multi-order fusion binary patterns are more effective for face representation than LBP. The reason may be that MFBP exploits extra higher-order derivatives which contain rich discriminant information.

Table 1. Compare recognition rate (%) of the proposed approach with other methods under the framework of Block-based FLD

Method	fb	fc	dupI	dupII	Average
Gabor+LBP+BFLD	99.0	98.9	87.3	75.6	90.2
Gabor+MFBP+BFLD	99.0	99.0	89.5	81.6	92.3
NSCT+LBP+BFLD	99.3	99.5	91.4	83.3	93.4
NSCT+MFBP+BFLD	99.5	100.0	92.9	88.9	95.3
NSCT+MFBP+BFLD+FSC (proposed method)	99.7	100.0	94.3	91.9	96.5

4.3 Evaluation of the Proposed Approach

Our proposed approach can be summarized that after extracting MFBP on NSCT coefficient images, we perform BFLD scheme and use FSC weights to combine all blocks to output the final score. In this section, we compare our method with several state-of-the-art approaches, which use the same standard FERET protocol. Results of comparison are displayed in Table 2. As can be seen from Table 2, our proposed framework achieves the highest accuracy on all four probe subsets and outperforms several state-of-the-art methods in literatures.

Table 2. Compare recognition rate (%) of the proposed approach with several state-of-the-arts methods

Method	fb	fc	dupI	dupII	Average
FERET97 best [11]	96.2	82.0	59.1	52.1	72.4
LBP [1]	97.0	79.0	66.0	64.0	76.5
LGBPHS [5]	98.0	97.0	74.0	71.0	85.0
HGPP_Weighted [3]	97.5	99.5	79.5	77.8	88.6
LNSCTBP+W_BKFLD [6]	99.0	99.0	92.0	86.0	94.0
EPFDA_LGBP [4]	99.6	99.0	92.0	88.9	94.8
Our Proposed Method	99.7	100.0	94.3	91.9	96.5

5 Conclusions

A new face recognition approach based on NSCT and MFBP has been proposed in this paper. Due to the ability of extracting both intrinsic geometrical information and texture information, NSCT can be used for effective face representation over the well-known Gabor features. On the other hand, MFBP which adds high-order derivatives can complete the original LBP operators. After calculating histograms of MFBP on NSCT decomposed images, BFLD feature reduction method and weight scheme based on Fisher separation criteria are chosen to further improve the performance. Experiments on FERET database show that the proposed framework could outperform many of the art-of-the-state approaches.

Acknowledgements. The work is partially supported by NSFC (No. 61101150), Shenzhen research fund (No. JCYJ20120831165730901).

References

1. 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)
2. Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local Derivative Pattern Versus Local Binary Pattern: Face Recognition with High-order Local Pattern Descriptor. *IEEE Transactions on Image Processing* 19(2), 533–544 (2010)
3. Zhang, B., Shan, S., Chen, X., Gao, W.: Histogram of Gabor Phase Patterns (HGPP): A Novel Object Representation Approach for Face Recognition. *IEEE Transactions on Image Processing* 16(1), 57–68 (2007)
4. Shan, S., Zhang, W., Su, Y., Chen, X., Gao, W.: Ensemble of Piecewise FDA Based on Spatial Histograms of Local (Gabor) Binary Patterns for Face Recognition. In: 18th International Conference on Pattern Recognition (ICPR), pp. 606–609 (2006)
5. 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: 10th IEEE International Conference on Computer Vision (ICCV), vol. 1, pp. 786–791 (2005)
6. Wang, B., Li, W., Liao, Q.: Face Recognition Based on Non-subsampled Contourlet Transform and Block-based Kernel Fisher Linear Discriminant. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1533–1536 (2012)
7. da Cunha, A.L., Zhou, J., Do, M.N.: The Non-subsampled Contourlet Transform: Theory, Design, and Applications. *IEEE Transactions on Image Processing* 15(10), 3089–3101 (2006)
8. Do, M.N., Vetterli, M.: The Contourlet Transform: An Efficient Directional Multi-resolution Image Representation. *IEEE Transactions on Image Processing* 14(12), 2091–2106 (2005)
9. Ojala, T., Pietikainen, M., Maenpaa, T.: Multi-resolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
10. Yan, K., Chen, Y., Zhang, D.: Gabor Surface Feature for Face Recognition. In: 1st Asian Conference on Pattern Recognition (ACPR), pp. 288–292 (2011)
11. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation methodology for Face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(10), 1090–1104 (2000)

Texture-Aware Fast Global Level Set Evolution

Souleymane Balla-Arabé¹, Xinbo Gao¹, and Lai Xu²

¹ School of Electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, P.R. China

² School of Design, Engineering and Computing, Bournemouth University, UK

balla_arabe_souleymane@ieee.org, xbgao@mail.xidian.edu.cn,
lxu@bournemouth.ac.uk

Abstract. Due to its intrinsic advantages such as the ability to automatically handle complex shapes and topological changes, the level set method has been widely used in image segmentation. Nevertheless, in addition to be computational expensive, it has the limitation to very often lead to a local minimum because of the energy functional to be minimized is non-convex. In this work, we use the geometric active contours and the image thresholding frameworks to design a novel method for global image segmentation. The local lattice Boltzmann method is used to solve the level set equation. The proposed algorithm is therefore effective and highly parallelizable. Experimental results on satellite, natural and medical images demonstrate the effectiveness and the efficiency of the proposed method when implemented using an NVIDIA graphics processing units.

Keywords: level set method, partial differential equations, graphics processing units, lattice Boltzmann method, image thresholding.

1 Introduction

The level set method (LSM) [1]-[3] is the geometric representation of active contours models [4]. It is a numerical technique for tracking interfaces and shapes. In recent years, it has attracted much more attention due to its advantageous intrinsic quality which allows to easily handle complex shapes and topological changes. In two-dimensional (2D) space, the LSM represents a closed curve in the plane as the zero level set of a three-dimensional (3D) curved surface. The curve has to move toward its interior or exterior normal until defining the boundary of the object of interest. The active contour evolution is driven by the level set equation (LSE) which is a partial differential equation and, in its general form can be expressed as

$$\frac{\partial \phi}{\partial t} = |\nabla \phi| (\alpha V + \beta \nabla \cdot \nabla \phi / |\nabla \phi|), \quad (1)$$

where ϕ is the level set function (LSF), $\nabla \phi$ is the gradient of ϕ , V is the speed function which drives the active contour towards the region boundaries, the second term in the brackets of the right hand represents the curvature and is used to smooth the contour, and α and β are user-controlling parameters. Generally two methods can be used to stop the evolving curve on the boundaries of the desired objects.

The first one uses an edge indicator depending on the gradient of the image like in classical snakes and active contour models [5]; and the second approach employs some regional statistics to stop the evolving curve [6]-[7]. The latter approach can detect objects which boundaries are not defined by clear edges and, comparing with the first one, it is more robust against noise since it uses regional information. The active contours without edges method proposed in [7] is one of the most widely used region-based LSM. In this model, the image is partitioned into two phases by the active contour; and the means of the intensity values inside and outside the contours are then used to decide if a given pixel is inside wherever outside the contour. The method is robust against noise but has the drawback to very often lead to a local minimum due to the fact that the energy functional to be minimized is non-convex.

In this paper, we propose a local region-based level set algorithm which combines simultaneously texture information, and the global but less accurate simple image thresholding result to constrain the active contour to detect the global minimum. Furthermore, we make the algorithm faster and suitable for parallel programming by using the lattice Boltzmann method (LBM) as an alternative approach to the LSE. The LBM is of recent use in image segmentation [8], [15] and [16]. It is second order accuracy both in time and space and can better handle the problem of time consuming because the non-linear term in the LSE, i.e., the curvature term, is implicitly computed. The LBM is firstly designed to simulate Navier-Stokes equations for an incompressible fluid [9]. The evolution equation of LBM is

$$f_i(\vec{r} + \vec{e}_i, t+1) - f_i(\vec{r}, t) = \Omega_{coll}, \quad (2)$$

where f_i is the particle distribution function and Ω_{coll} is, in this paper, the Bhatnager-Gross-Krook (BGK) collision model with a body force \vec{F} .

$$\Omega_{coll} = [1/\tau] \cdot [f_i^{eq}(\vec{r}, t) - f_i(\vec{r}, t)] + [D/bc^2] \cdot \vec{F} \cdot \vec{e}_i, \quad (3)$$

where D is the grid dimension, b is the link at each grid point, c is the length of each link which is set to 1 in this paper, τ represents the relaxation time and f_i^{eq} the local Maxwell-Boltzmann equilibrium particle distribution function which can be expressed in discrete form as follows when modeling typical diffusion phenomenon,

$$f_i^{eq}(\rho) = \rho A_i \text{ with } \rho = \sum_i f_i, \quad (4)$$

where ρ is the macroscopic fluid density. By performing the Chapman-Enskog expansion the following diffusion equation can be recovered from LBM [9],

$$\partial\rho/\partial t = \beta \operatorname{div}(\nabla\rho) + F. \quad (5)$$

Substituting ρ by the signed distance function ϕ in Eq. (5), the LSE can be recovered. In our model we use the D2Q5 ($D = 2, d = 5$) LBM lattice structure. The body force F acts as the image data link for the LBM solver. Since the LBM is local and thus suitable for parallel programming, the proposed algorithm is accelerated using an NVIDIA graphics processing units (GPU). Furthermore, due to the use of region and texture information, it is robust against noise and effective when detecting objects for which boundaries are not clearly defined.

The remainder of this paper is organized as follows. Section 2 details the proposed image segmentation model. Experimental results are presented in Section 3, whereas the conclusions are drawn in Section 4.

2 The Proposed Level Set Method

This section describes the conception of the proposed fast convex level set algorithm. The energy functional that we propose to minimize is designed as

$$\mathcal{E}(\phi) = \lambda \mathcal{E}_{\text{texture}}(\phi) + \beta \mathcal{E}_{\text{thresh}}(\phi) + \gamma \mathcal{E}_{\text{reg}}(\phi), \quad (6)$$

where ϕ is the level set function. λ , β and γ are positive controlling parameters. The energy term $\mathcal{E}_{\text{texture}}(\phi)$ is texture based, and is defined as follows with Ω the image domain, Ω_{in} and Ω_{out} the image domain respectively inside and outside the evolving curve,

$$\mathcal{E}_{\text{texture}}(\phi) = (1/2) \int_{\Omega_{in}} (\theta(\psi) - \theta(m_{in}))^2 dx + (1/2) \int_{\Omega_{out}} (\theta(\psi) - \theta(m_{out}))^2 dx, \quad (7)$$

where $\psi(x) = [I(x), I_f(x), s(x)] \in R^3$, in which $I(x) \in R$ is the intensity of pixel x . The 2-tuple $[I_f(x), s(x)] \in R^2$ is a simple descriptor of the texture information at pixel x , where $I_f(x)$ is the filtered intensity of pixel x and $s(x)$ is the standard variance of the intensities of the pixels in the neighborhood of pixel x . θ is a transformation function defined by the Gaussian kernel

$$k(x, y) = \langle \theta(x), \theta(y) \rangle = \exp \left[-\| [I_f(x), s(x)] - [I_f(y), s(y)] \|_2^2 / \sigma^2 \right], \quad (8)$$

where σ is the adjustable parameter. We can thus rewrite Eq. (7) as

$$\begin{aligned} \mathcal{E}_{\text{texture}}(\phi) &= [1/2] \int_{\Omega_{in}} (\theta(\psi) - \theta(m_{in}))^2 dx + [1/2] \int_{\Omega_{out}} (\theta(\psi) - \theta(m_{out}))^2 dx \\ &= [1/2] \int_{\Omega_{in}} (\theta(\psi)^2 + \theta(m_{in})^2 - 2\theta(\psi)\theta(m_{in})) dx \\ &\quad + [1/2] \int_{\Omega_{out}} (\theta(\psi)^2 + \theta(m_{out})^2 - 2\theta(\psi)\theta(m_{out})) dx \\ &= [1/2] \int_{\Omega_{in}} (k(\psi, \psi) + k(m_{in}, m_{in}) - 2k(\psi, m_{in})) dx \\ &\quad + [1/2] \int_{\Omega_{out}} (k(\psi, \psi) + k(m_{out}, m_{out}) - 2k(\psi, m_{out})) dx, \end{aligned} \quad (9)$$

from Eq. (8), we have $k(\psi, \psi) = 1$, $k(m_{in}, m_{in}) = 1$ and $k(m_{out}, m_{out}) = 1$. Thus Eq. (9) can be simplified as

$$\mathcal{E}_{\text{texture}}(\phi) = \int_{\Omega} (1 - k(\psi, m_{in})) H(\phi) + (1 - k(\psi, m_{out})) (1 - H(\phi)) dx, \quad (10)$$

where H is the Heaviside function. m_{in} and m_{out} are respectively the local mean values inside and outside the evolving contour. They are defined as

$$m_{out}(x) = \int_{\Omega} k(x, y) I(y) (1 - H(\phi)) dy / \int_{\Omega} k(x, y) (1 - H(\phi)) dy, \quad (11)$$

$$m_{in}(x) = \int_{\Omega} k(x, y) I(y).H(\phi) dy / \int_{\Omega} k(x, y) H(\phi) dy, \text{ with } k(x, y) = \begin{cases} 1, & |x - y| < r \\ 0, & \text{otherwise} \end{cases}, \quad (12)$$

where x and y are spatial variables, and r a radius constant.

The energy term $\mathcal{E}_{thresh}(\phi)$ is based on the result obtained after a simple thresholding of the image that we want to segment. In this paper we use the Otsu's thresholding method, although any kind of simple thresholding method can be used. $\mathcal{E}_{thresh}(\phi)$ is designed as

$$\begin{aligned} \mathcal{E}_{thresh}(\phi) &= [1/2].(1 - sign(\phi_{thresh})sign(\phi)) \\ &= [1/2].(1 - (2H(\phi_{thresh}) - 1)(2H(\phi) - 1)) \\ &= H(\phi_{thresh}) + H(\phi)(1 - H(\phi_{thresh})), \end{aligned} \quad (13)$$

where ϕ_{thresh} is a signed distance function obtained from the thresholded image. It is positive in one class and negative in the other one.

The regularization term $\mathcal{E}_{reg}(\phi)$ is used as a constraint on the evolving contour, and can be expressed as in [12]

$$\mathcal{E}_{reg}(\phi) = \int_{\Omega} |\nabla H(\phi)| dx. \quad (14)$$

The proposed energy functional (Eq. (6)) can therefore be rewritten as

$$\begin{aligned} \mathcal{E}(\phi) &= \lambda \int_{\Omega} (1 - k(\psi, m_{in})) H(\phi) + (1 - k(\psi, m_{out})) (1 - H(\phi)) dx \\ &\quad + \beta (H(\phi_{thresh}) + H(\phi)(1 - H(\phi_{thresh}))) + \gamma \int_{\Omega} |\nabla H(\phi)| dx, \end{aligned} \quad (15)$$

By using the gradient descent method

$$\partial \phi / \partial t = -\partial \mathcal{E} / \partial \phi, \quad (16)$$

where $\partial \mathcal{E} / \partial \phi$ is the Gâteaux derivative [12] of \mathcal{E} , we obtain the following LSE

$$\partial \phi / \partial t = \delta(\phi)[\lambda(k(\psi, m_{in}) - k(\psi, m_{out})) + \beta(1 - H(\phi_{thresh})) + \gamma \operatorname{div}(\nabla \phi / |\nabla \phi|)]. \quad (17)$$

The gradient projection method allows us to replace $\delta(\phi)$ by $|\nabla \phi|$, and as ϕ is a signed distance function we have $|\nabla \phi| = 1$. Thus Eq. (17) can be rewritten as

$$\partial \phi / \partial t = \lambda(k(\psi, m_{in}) - k(\psi, m_{out})) + \beta(1 - H(\phi_{thresh})) + \gamma \operatorname{div}(\nabla \phi / |\nabla \phi|), \quad (18)$$

which is similar to Eq. (5) with the body force expressed as

$$F = \lambda(k(\psi, m_{in}) - k(\psi, m_{out})) + \beta(1 - H(\phi_{thresh})). \quad (19)$$

The proposed level set equation can therefore be solved using the following lattice Boltzmann evolution equation

$$\begin{aligned} f_i(\vec{r} + \vec{e}_i, t+1) &= f_i(\vec{r}, t) + [1/\tau].[\bar{f}_i(\vec{r}, t) - f_i(\vec{r}, t)] + [D/bc^2].(\lambda(k(\psi, m_{in}) - k(\psi, m_{out}))) \\ &\quad + \beta(1 - H(\phi_{thresh}))), \end{aligned} \quad (20)$$

without the necessity of explicitly compute the curvature since it is implicitly handled by the LBM.

The principal implementation steps of the proposed method are as follows:

<i>Steps</i>	<i>Instructions</i>
1	Segment the given image to be processed with thresholding method, find the ϕ_{thresh} and initialize ϕ as $\phi = \phi_{thresh}$.
2	Compute the body force F with Eq. (19).
3	Resolve the LSE using LBM with Eq. (18).
4	Accumulate the $f_i(\vec{r}, t)$ values at each grid point with eq. (4), which generates updated values of ϕ and find the contour.
5	If the evolving curve has not converged go back to step 2.

We implemented the body force and only the collision step on GPU, since this step is fully local the speed up is considerable, the streaming step is executed on the CPU. The optimized Matlab function *arrayfun* is used to execute the code on the GPU.

3 Experiments and Analysis

In this section, we carried out several experiments using various kinds of images in order to demonstrate the performance of the proposed method. The proposed method is compared subjectively and objectively, in terms of efficiency, speed and effectiveness, with three level set methods. The first one is an edge-based method proposed by Li *et al.* in [2] where the main idea was to perform a level set segmentation without re-initialization. The second method is the well known global region-based level set segmentation method proposed by Chan and Vese (C-V) in [7]. The third one is a lattice Boltzmann based method proposed by Hagan *et al.* in method [3]. For the objective evaluation, we use the Zeboudj's contrast [13] and the Rosenberger's criterion [14] as metrics. The better is the segmentation result; the higher are the two criteria. In all the experimental results, the interior of the final level set function is represented by black pixels, and the exterior by white pixels. We implemented the method using the parallel computing toolbox of Matlab R2012a installed on a PC AMD Athlon (TM) 5200 processor with a clock speed of 2.31 GHz, 2 GB of RAM and possessing the NVIDIA GPU GT 430. We fixed $r = 3$, $\beta = 0.1$ and $\lambda = 1$.

In Fig.1, we apply our algorithm on a real world image of rabbit; Fig.1 (b) shows the result after the Otsu's thresholding, Fig.1 (c) and Fig.1 (d) show the promising final result.

From Fig.2 to Fig.4, we demonstrate the ability of the proposed image segmentation on natural images. In Fig.4, we use a medical image and in Fig.5, we use a very high resolution satellite image. The executive times and the results of the objective evaluation are displayed by TABLES 1 to 5.

When visually analyzing the experimental results, it can be subjectively noticed that the proposed method gives the most promising results in term of global segmentation. This is objectively demonstrated by the fact that, in all the experiments, it has almost the highest Zeboudj's and Rosenberger's measures. The C-V's method is easily trapped in a local minimum, and fall to give the global segmentation results like in Fig.2, Fig.3 and Fig.4. The Hagan's method is not efficient in presence of texture

and can give an over segmented result like in Fig.3. The Li's method cannot detect objects without strong edges, this result in some leakages on the boundaries, and the method detects local minimum.

Furthermore, the proposed algorithm is very interesting in term of executive time. It is faster than the C-V and the Li's method. Only the Hagan's method has its same order of speed.

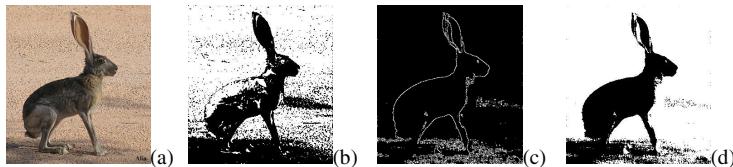


Fig. 1. (a): original image. (b): Otsu's thresholding. (c): proposed method: final contour. (d): proposed method: black pixels represent the interior of the final contour and the white pixels the exterior.

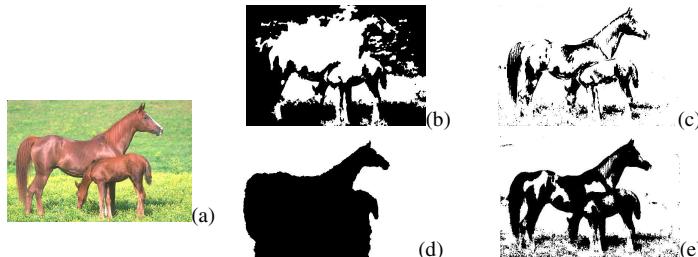


Fig. 2. (a): original image. (b): C-V method. (c): Hagan's algorithm. (d): Li's algorithm (e): segmentation result of the proposed method.

Table 1. Comparison of Executive Times and Objective Evaluations

Methods	Our method	C-V	Hagan	Li
Time(s)	0.734	225.123	1.687	233.86
Zeboudj	0.87	0.63	0.54	0.37
Rosenberger	0.90	0.69	0.46	0.50

NB: Image dimensions 481 X 321

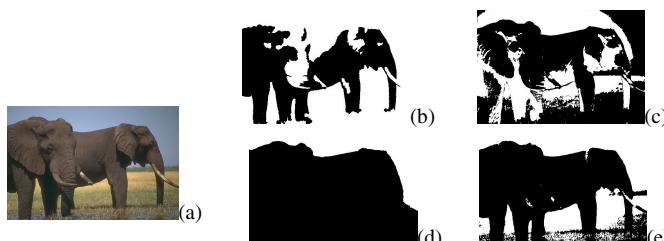


Fig. 3. (a): original image. (b): C-V method. (c): Hagan's algorithm. (d): Li's algorithm (e): segmentation result of the proposed method.

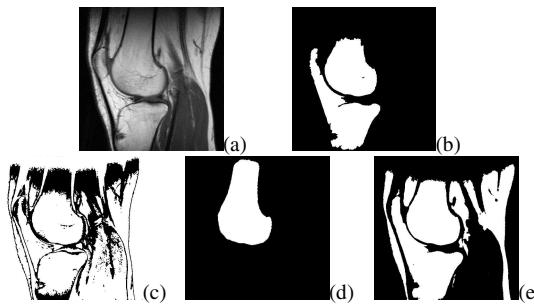


Fig. 4. (a): original image. (b): C-V method. (c): Hagan's algorithm. (d): Li's algorithm (e): segmentation result of the proposed method.

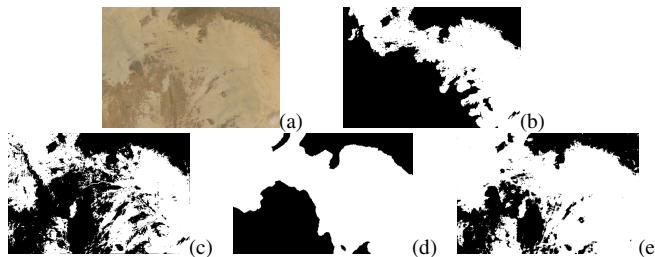


Fig. 5. (a): original image. (b): C-V method. (c): Hagan's algorithm. (d): Li's algorithm (e): segmentation result of the proposed method.

Table 2. Comparison of Executive Times and Objective Evaluations

Methods	Our method	C-V	Hagan	Li
Time(s)	0.43	107.453	0.58	90.067
Zeboudj	0.77	0.69	0.45	0.42
Rosenberger	0.82	0.71	0.59	0.63

NB: Image dimensions 481 X 321.

Table 3. Comparison of Executive Times and Objective Evaluations

Methods	Our method	C-V	Hagan	Li
Time(s)	0.736	114.79	1.526	41.176
Zeboudj	0.58	0.48	0.52	0.51
Rosenberger	0.86	0.73	0.70	0.75

NB: Image dimensions 512 X 512.

Table 4. Comparison of Executive Times and Objective Evaluations

Methods	Our method	C-V	Hagan	Li
Time(s)	1.003	239.716	2.352	131.92
Zeboudj	0.87	0.54	0.65	0.43
Rosenberger	0.91	0.72	0.62	0.54

NB: Image dimensions 1323 X 888

4 Conclusion

We have presented a method which combines the advantages of the LSM and the image thresholding technique in term of global segmentation. This allows it to easily handle complex shapes, and simultaneously guaranty its convergence towards the global minimum. The incorporation of texture information in the segmentation process, via some kernel-based techniques, increases the ability of the proposed algorithm. Furthermore, a GPU-based acceleration makes it interesting in term of executive time. Experiments have demonstrated subjectively and objectively the good performance of the proposed model.

References

- [1] Balla-Arabé, S., Gao, X.: A Multiphase Entropy-Based Level Set Algorithm for MR Breast Image Segmentation Using Lattice Boltzmann Model. In: Yang, J., Fang, F., Sun, C. (eds.) IScIDE 2012. LNCS, vol. 7751, pp. 8–16. Springer, Heidelberg (2013)
- [2] Li, C., Xu, C., Gui, C., Fox, M.: Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on Image Processing* 19(12), 3243–3254 (2010)
- [3] Hagan, A., Zhao, Y.: Parallel 3D image segmentation of large data sets on a GPU cluster. In: Bebis, G., et al. (eds.) ISVC 2009, Part II. LNCS, vol. 5876, pp. 960–969. Springer, Heidelberg (2009)
- [4] Zhu, S., Yuille, A.: Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(9), 884–900 (1996)
- [5] Gao, X.-B., Wang, B., Tao, D., Li, X.: A relay level set method for automatic image segmentation. *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics* 41(2), 518–525 (2011)
- [6] Balla-Arabé, S., Gao, X.-B., Wang, B.: GPU Accelerated Edge-Region Based Level Set Evolution Constrained By 2D Gray-scale Histogram. *IEEE Transactions on Image Processing* 22(7), 2688–2698 (2013)
- [7] Chan, T., Vese, L.: Active contours without edges. *IEEE Transactions on Image Processing* 10(2), 266–277 (2001)
- [8] Balla-Arabé, S., Gao, X.-B., Wang, B.: A Fast and Robust Level Set Method for Image Segmentation Using Fuzzy Clustering and Lattice Boltzmann Method. *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics* 43(3), 910–920 (2013)
- [9] Zhao, Y.: Lattice Boltzmann based PDE solver on the GPU. *The Visual Computer* 24(5), 323–333 (2007)
- [10] Bhatnager, P., Gross, E., Krook, M.: Phys. Rev. 94, 511 (1954)
- [11] Buick, J., Created, C.: Gravity in a lattice Boltzmann model. *Phys. Rev. E* 61(5), 5307–5320 (2000)
- [12] Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. CRC Press, Boca Raton (1992)

- [13] Zeboudj, R.: Filtrage, seuillage automatique, contraste et contours: du pré-traitement à l'analyse d'images. PhD thesis, Saint Etienne University (1988)
- [14] Rosenberger, C., Chehdi, K.: "Genetic fusion: Application to multi-components image segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2219–2222 (2000)
- [15] Balla-Arabé, S., Gao, X.-B.: Image multi-thresholding by combining the lattice Boltzmann model and a localized level set algorithm. Neurocomputing 93, 106–114 (2012)
- [16] Balla-Arabé, S., Wang, B., Gao, X.-B.: Level set region based image segmentation using lattice Boltzmann method. In: Proc. 7th Int. Conf. Computational Intelligence and Security, Sanya, China, pp. 1159–1163 (2011)

A Novel Metric for Image Denoising Algorithms

Yingtao Zhang¹, H.D. Cheng², Jianhua Huang¹, and XiangLong Tang¹

¹ School of Computer Science and Technology, Harbin Institute of Technology, China

² Dept. of Computer Science, Utah State University, Logan, UT 84322 U.S.A

Abstract. Denoising algorithms, especially, the ones with contrast enhancement capability have many important applications.

However, there is no an effective and accurate measurement for evaluating their performance objectively. This introduces a new metric, *HME* (Homogeneity Mean Error), to assess the denoising algorithms, especially those with enhancement capability. *HME* is based on the homogeneity property of each pixel which is sensitive to the changes of the structural information and noise levels, but insensitive to the changes of the contrast. Therefore, it can be utilized for evaluating the denoising algorithms. Various experiments are performed on images corrupted with different type of noise, the results demonstrate that *HME* is an effective and accurate metric for assessing the denoising algorithms with/without contrast enhancement.

Keywords: objective criterion, denoising performance, *HMD* (Homogeneity Mean Difference), contrast enhancement, image quality assessment (QA).

1 Introduction

Many denoising algorithms were developed to reduce the noise and enhance the details of the images, which have been applied to natural images, synthetic aperture radar (SAR) images, medical images, microarray images, etc [1-4]. However, there is no an effective and accurate criterion for assessing the performance of denoising algorithms, especially, those have enhancement capability. In the last three decades, many efforts have dedicated to develop quality assessments [5]. Quantitative evaluation of the algorithms focuses on several criteria such as edge preservation, intrinsic texture preservation, mean preservation, variance reduction in a homogeneous region, etc [6].

Most of metrics are the full-reference quality assessment, i.e., a complete reference image is known. There are two kinds of full-reference metrics. One type is the error sensitive measure, which considers the distorted signal as the sum of a reference signal and an error signal. The other type is the structural similarity measure, which is based on the hypothesis that the HVS (human vision system) is highly adaptive for extracting structural information. One recently proposed approach to image fidelity measurement is the structural similarity (*SSIM*) index [5].

We propose a novel objective criterion, *HME* (Homogeneity Mean Error), which can assess the performance of denoising algorithms effectively and accurately. Homogeneity is largely related to the local information of an image and reflects how uniform a region is [7]. There are a few definitions of the homogeneity [8-11] for different applications. Here, we define the homogeneity based on the edge value,

standard deviation, and entropy of each pixel. The homogeneity mean can be calculated. Then, HME , the difference between the homogeneity means of the processed image and reference image, is computed.

2 Image Quality Measurements

Traditional metrics measure the denoise performance by comparing the distance between the reference image and processed image [12]. However, they do not correlate with the perceived quality well, and their limitations have been widely recognized [13]. Another recently proposed class of metrics is based on Structural SIMilarity (SSIM) [5]. Such measure takes into account point-by-point distortions that may not be relevant to the perception quality either.

2.1 Error Sensitivity Metrics

The most commonly used objective fidelity measure is MSE which is defined as:

$$MSE = \frac{1}{M \cdot N} \sum_{(i,j=1)}^{M \cdot N} (\hat{I}(i, j) - I(i, j))^2 \quad (1)$$

where I and \hat{I} are the reference and filtered images of size $M \cdot N$, respectively. It can easily find that some filtered images with the same MSE may have quite different errors, and some of them are much more visible than others.

MSE is also often converted into a peak signal-to-noise ratio ($PSNR$) measure [14]:

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \quad (2)$$

where L is the dynamic range of allowable pixel intensities. For gray level images, if the intensity level has 8 bits, then $L = 255$. The $PSNR$ is useful when the images having different dynamic intensity ranges are compared, otherwise, it contains no new information than MSE . Since all image pixels are treated equally in MSE , the content-dependent variations of the images cannot be evaluated well, and the important feature and structure information are ignored.

SNR is defined as the ratio of the average signal power to the average noise power. SNR is also used by many denoising algorithms, which assumes the distortion caused only by the additive noise. The SNR of an image with size $M \cdot N$ can be defined as:

$$SNR = 10 \log_{10} \frac{\sum_{(i,j=1)}^{M \cdot N} (I(i, j) - \text{mean}(I(i, j)))^2}{\sum_{(i,j=1)}^{M \cdot N} (\hat{I}(i, j) - I(i, j))^2} \quad (3)$$

where $\text{mean}(\cdot)$ is the mean of the intensities of the image.

However, the correlation between SNR and visual quality is very poor [15].

Contrast enhancement can make all of the above measures of an enhanced image worse, even the enhanced image is much better visually and perceptually.

Another measure to compare edge preservation capability is “figure of merit” [16]:

$$FOM = \frac{1}{\max\{\hat{N}, N_{ideal}\}} \sum_{i=1}^{\hat{N}} \frac{1}{1 + d_i^2 \alpha} \quad (4)$$

where N_{ideal} and \hat{N} are the numbers of the ideal and detected edge pixels, respectively; d_i is the Euclidean distance between the i th detected edge pixel and the nearest ideal edge pixel; and α , a constant, is typically set to 1/9. FOM ranges between 0 and 1, with unity for the ideal edge detection. FOM strongly depends on the method to compute the edge map, i.e., different edge detectors could generate different FOM values [6].

2.2 Structural Similarity Measures

The basic $SSIM$ index is a real number in the range [-1, 1], and is calculated using the second order statistics of the reference and the distorted images:

$$SSIM(x, y) = \frac{(2\bar{x}\bar{y} + C_1)(2\delta_{xy} + C_2)}{(\bar{x}^2 + \bar{y}^2 + C_1)(\delta_x^2 + \delta_y^2 + C_2)} \quad (5)$$

where x and y are two nonnegative image signals, \bar{x} and \bar{y} are the mean intensities, δ_x^2 and δ_y^2 are the variances, and σ_{xy} is the covariance of x and y . C_1 and C_2 are the small real constants to avoid instability. Other forms of such metric can be found in [5]. Although $SSIM$ metrics have introduced a way to investigate the image fidelity, they also have limitations.

The metrics discussed above cannot evaluate the performance of the filters, especially, those with enhancement capability.

3 HME (Homogeneity Mean Error)

The fundamental idea behind our approach is that we want to propose a metric which is not sensitive to the enhancement, but sensitive to the noise level. The homogeneity mean error (*HME*) just is such a metric.

Suppose that g_{ij} is the gray level of pixel (i, j) in an image I of size $M \times N$, and w_{ij} is a window of size $d \times d$ centered at (i, j) for computing *edge_value*, *entropy*, and *standard_deviation*. Here, $d = 3$.

Sobel operator is used to calculate the edge value:

$$e_{ij} = \sqrt{s_{1_{ij}}^2 + s_{2_{ij}}^2} \quad (6)$$

where s_1 and s_2 correspond to the results from the row mask and column mask, respectively.

Then standard deviation is:

$$Sd_{ij} = \sqrt{\frac{1}{d^2} \sum_{p=i-\frac{d-1}{2}}^{i+\frac{d-1}{2}} \sum_{q=j-\frac{d-1}{2}}^{j+\frac{d-1}{2}} (g_{pq} - m_{ij})^2} \quad (7)$$

where $0 \leq i \leq M - 1$ and $0 \leq j \leq N - 1$ and m_{ij} is the mean intensity value of the window, whose value is:

$$m_{ij} = \frac{1}{d^2} \sum_{p=i-\frac{d-1}{2}}^{i+\frac{d-1}{2}} \sum_{q=j-\frac{d-1}{2}}^{j+\frac{d-1}{2}} g_{pq} \quad (8)$$

The entropy of a pixel can be calculated as:

$$H_{ij}(p_i) = -\frac{1}{d^2} \sum p_i \log_2 p_i, \quad i=1, \dots, n \quad (9)$$

where i represents the index of the intensity level, p_i denotes the probability of the i th intensity, and n is the number of intensity levels in the window.

Normalize the edge value, standard deviation, and entropy to achieve the computation consistency:

$$E_{ij} = \frac{e_{ij}}{e_{\max}}, V_{ij} = \frac{v_{ij}}{v_{\max}}, H_{ij} = \frac{h_{ij}}{h_{\max}} \quad (10)$$

where $e_{\max} = \max\{e_{ij}\}$, $v_{\max} = \max\{v_{ij}\}$, $h_{\max} = \max\{h_{ij}\}$.

Let HO_{ij} denote the homogeneity value of pixel (i, j) :

$$\begin{aligned} HO_{ij} &= \overline{E_{ij}} \cdot \overline{V_{ij}} \cdot \overline{H_{ij}} \\ &= (1 - E_{ij}) \cdot (1 - V_{ij}) \cdot (1 - H_{ij}) \end{aligned} \quad (11)$$

Finally, HME can be obtained:

$$HME(R, P) = \frac{1}{M \cdot N} \sum_{(i,j)=1}^{M \cdot N} |HO_{ij}^R - HO_{ij}^P| \quad (12)$$

where HO_{ij}^P and HO_{ij}^R denote the homogeneity values of the pixels (i, j) in the processed image and reference image, respectively.

4 Characteristics of HME

4.1 Noise Sensitivity of HME

Three types of noises (Gaussian, salt and pepper, and speckle noises) were used to demonstrate HME 's effectiveness and usefulness in assessing the noise removing capability. A variety of images have been tested, due to page limit, only three images are presented here. In Fig.1, we show the results of applying speckle noise with different standard deviation u to a synthetic image. The speckle noise model in [17] was used. As we anticipated, HME changes with noise level proportionally. When u is 0.05, HME is 0.36. If noise level is high enough, HME is close to 1, as shown in Fig.1(c). In Figs. 2 and 3, we show the results by applying Gaussian white noise and salt and pepper noise with different levels to real images.

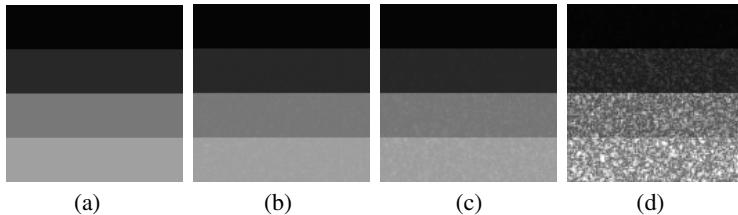


Fig. 1. HME values of the images with different levels of speckle noise (a) The original image (b) Corrupted image with $u = 0.05$, and $HME = 0.36$ (c) Corrupted image with $u = 0.1$, and $HME = 0.48$. (d) Corrupted image with $u = 1.0$, and $HME = 0.81$.

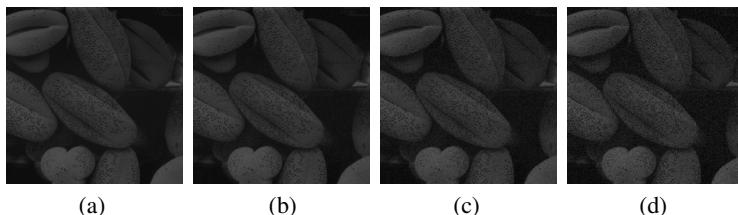


Fig. 2. HME values of the images with different levels of additive Gaussian white noise (a) The original Lena image of size 512×512 , (b) Corrupted image with $\sigma = 5$, $HME = 0.25$ (c) Corrupted image with $\sigma = 10$, $HME = 0.28$ (d) Corrupted image with $\sigma = 20$, $HME = 0.29$

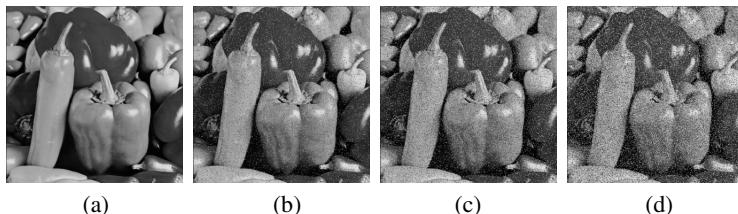


Fig. 3. HME values of the images with different levels of salt and pepper noise (a) The original Lena image of size 353×353 , (b) Corrupted image with 5% salt and pepper noise, $HME = 0.03$ (c) Corrupted image with 10% salt and pepper noise, $HME = 0.05$ (d) Corrupted image with 20% salt and pepper noise, $HME = 0.08$

From Figs. 1-3, we can conclude that *HME* is very sensitive to different kinds of noises of different levels, therefore, it can be used for evaluating the noise removal capability of the denoising filters.

4.2 Insensitivity of *HME* to Enhancement

Fig. 4 shows an example of the images with 8 different level contrasts from 0.091 to 0.939. The contrast is calculated as [18]:

$$C = \frac{f - b}{f + b} \quad (13)$$

where f and b are the maximum and minimum gray values, respectively, in the image. All their HME values are zero as shown in Table 1. The results demonstrate that HME is insensitive to the enhancement.

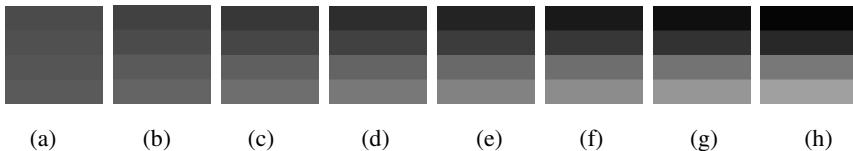


Fig. 4. Images with different contrasts

Table 1. HME values of the images of Fig.4

image	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
C	0.091	0.212	0.333	0.455	0.576	0.697	0.818	0.939
HME	0	0	0	0	0	0	0	0

4.3 Comparing HME with the Existing Metrics

In the experiments, we compare HME with the commonly used metrics: MSE , SNR and $SSIM$. We employed the S-function to enhance the images, since it is very popular and easy to implement.

Fig.5 shows the test images having additive Gaussian white noises and the corresponding enhanced images by the S-function. The results of applying different metrics to the images in Fig. 5 are listed in Table 2.



Fig. 5. In each column, the images with the same level of additive Gaussian white noise. From left to right columns, the noise levels are: $\sigma = 0$, $\sigma = 5$, $\sigma = 10$ and $\sigma = 15$. The images in the first row are the noise-corrupted, and the images in the second row are the noise-corrupted and enhanced with the S-function.

Table 2. Performance comparison of metrics on Fig. 5

	MSE				SNR			
σ	0	5	10	15	0	5	10	15
No Enhancement	0	84	104	212	∞	15.42	13.51	10.46
Enhancement	667	682	652	811	5.39	5.38	5.46	5.43
	SSIM				HME			
σ	0	5	10	15	0	5	10	15
No Enhancement	1	0.812	0.611	0.485	0	0.109	0.127	0.135
Enhancement	0.856	0.684	0.476	0.348	0.038	0.098	0.119	0.131

From Table 2, we can observe the following facts:

1. *MSE* and *SNR* are very sensitive to enhancement which makes *MSE* higher and *SNR* lower. It implies that enhancement makes *MSE* and *SNR* worse, even the corresponding images are better visually and perceptually.
2. *SSIM* is also sensitive to enhancement. For example, the value changes from 1 to 0.856 in the first column of *SSIM*, corresponding to the original and enhanced images. In addition, the value of *SSIM* cannot demonstrate the noise levels of the enhanced images well. For instance, *SSIM* of Fig. 5(h) with $\sigma=10$ is lower than *SSIM* of Fig. 5 (d) with $\sigma=15$.
3. *HME* is insensitive to enhancement. The difference between the values in the same column (Table 2) is very small. Furthermore, the value is lower than any one in the corresponding right columns, i.e., it is sensitive to noise level. Hence, it can be used as an effective and objective criterion to measure the performance of the filters with /without enhancement.

5 Conclusions

In this paper, we propose a novel objective criterion for evaluating the performance of denoising filters. The proposed *HME* is sensitive to noise and insensitive to enhancement, therefore, it can be used for assessing the performance of the denoising filters, especially, of the denoising filters with enhancement while the existing metrics cannot solve such tasks well.

Acknowledgments. This work is supported, in part, by National Science Foundation of China; the Grant numbers is 61100097.

References

1. Zhang, B., Allebach, J.P.: Adaptive bilateral filter for sharpness enhancement and noise removal. *IEEE Trans. Image Processing* 17(5), 664–678 (2008)
2. Xie, J., Jiang, Y.F., Tsui, H.T., Heng, P.A.: Boundary enhancement and speckle reduction for ultrasound images via salient structure extraction. *IEEE Trans. Biomedical Engineering* 53(11), 2300–2309 (2006)
3. Dong, Y., Milne, A.K., Forster, B.C.: Toward edge sharpening: a SAR speckle filtering algorithm. *IEEE Trans. Geoscience and Remote Sensing* 39(4), 851–863 (2001)
4. Wang, X.H., Istepanian, R.S.H., Song, Y.H.: Microarray Image Enhancement by Denoising Using Stationary Wavelet Transform. *IEEE Trans. Nanobioscience* 2(4), 184–189 (2003)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing* 13(4), 600–612 (2004)
6. Yu, Y., Acton, S.T.: Speckle reducing anisotropic diffusion. *IEEE Trans. Image Processing* 11(11), 1260–1270 (2002)
7. Gonzalez, R.C., Wintz, P.: Digital image processing. Addison-Wesley (1987)
8. Cheng, H.D., Li, J.G.: Fuzzy homogeneity and scale space approach to color image segmentation. *Pattern Recognition* 35, 373–393 (2002)
9. Cheng, H.D., Xue, M., Shi, X.J.: Contrast enhancement based on a novel homogeneity measurement. *Pattern Recognition* 36, 2687–2697 (2003)
10. Cheng, H.D., Chen, C.H., Chiu, H.H., Xu, H.J.: Fuzzy homogeneity approach to multilevel thresholding. *IEEE Trans. Image Processing* 7(7), 1084–1088 (1998)
11. Pok, G., Liu, J.-C., Nair, A.S.: Selective removal of impulse noise based on homogeneity level information. *IEEE Trans. Image Processing* 12(1), 85–92 (2003)
12. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Processing* 14(12), 1217–1228 (2005)
13. Bae, S.H., Pappas, T.N., Juang, B.H.: Subjective evaluation of spatial resolution and quantization noise tradeoffs. *IEEE Trans. Image Processing* 18(3), 495–508 (2009)
14. Wang, Z., Bovik, A.C.: Mean Squared Error: Leave It or Love It? *IEEE Signal Processing Magazine*, 98–117 (January 2009)
15. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. *IEEE Trans. Image Processing* 9(4), 636–650 (2000)
16. Pratt, W.K.: Digital image processing. Wiley, New York (1977)
17. Yue, Y., Croitoru, M.M., Bidani, A., Zwischenberger, J.B., Clark Jr., J.W.: Nonlinear multiscale wavelet diffusion for speckle suppression and edge enhancement in ultrasound images. *IEEE Trans. Medical Imaging* 25(3), 297–311 (2006)
18. Morrow, W.M., Paranjape, R.B., Rangayyan, R.M., Desautels, J.E.L.: Region-based contrast enhancement of mammograms. *IEEE Trans. Medical Imaging* 11, 392–406 (1992)

Adaptive Weight Optimization for Classification of Imbalanced Data

Wenhai Huang¹, Guojie Song^{1,*}, Man Li², Weisong Hu², and Kunqing Xie¹

¹ Key Laboratory of Machine Perception, Ministry of Education
Peking University, Beijing, 100871, China

² NEC Labs, China
gjsong@pku.edu.cn

Abstract. One popular approach for imbalance learning is weighting samples in rare classes with high cost and then applying cost-sensitive learning methods to deal with imbalance in classes. Weight of a class is usually determined by proportion of samples in each class in training set. This paper analyzes that sample proportions of training set and testing set may vary in some range and it would compromise performance of learned classifier. This problem becomes serious when class distribution is extremely high imbalanced. Based on the analysis, an adaptive weighting approach aiming at finding a group of proper weights for classes is proposed. We employ evolutionary algorithm to optimize weight configuration to ensure overall performance of classifier in both training set and possible testing sets. Experimental results on a wide variety of datasets demonstrate that our approach could achieve better performances.

1 Introduction

Classification in the presence of class imbalance is a pervasive problem in a wide variety of real-world areas such as medicine, biology, finance and internet. When number of instances in one class overwhelms other classes, many traditional classification methods (e.g. decision tree, SVM) assuming or expecting balanced class distribution would fail at providing favorable accuracies across all classes. Classifier tends to be biased on preferring majority classes.

Existing solutions to class imbalance problem could be divided into two categories: data level approaches and algorithm level approaches[4,12]. At the data level, these methods focus on re-sampling training set to make classes balanced. It over-samples minority classes[2] or under-samples[6] majority classes to balance class distribution. At the algorithm level, these solutions assign larger weights or costs to minority classes and then apply cost-sensitive learning methods to cope with class imbalance problem[10]. Many researches are focusing on how to make standard classification methods cost sensitive[7,11].

Assigning larger weights to minority classes could increase importance of minority samples so that they cannot be ignored easily. Weight or cost of each class

* Corresponding author.

is usually difficult to choose without prior knowledge. A natural and common method of determining weight distribution is using proportion of classes in training set[5,9,13]. Despite its nice theoretical properties, this choice of weights may have problems for practice. Instance distribution of real data set and training set may vary a lot even when training set is random-sampled[3]. Weighting by proportion of training set could only ensure best performance on training set.

From observation and theoretical analysis, we find that distribution difference would become larger when data becoming more imbalanced. Based on the analysis, an adaptive weighting mechanism is proposed. Our objective is to find an optimal weight distribution which could lead good performance not only on training set but also on various possible testing sets. We first sample several testing sets with different class distributions from training set to simulate possible distributions of real data set. Then we employ an Evolutionary Algorithm to search for optimal weight distribution. Overall performance of classifier on previous sampled testing sets are used to guide optimization process. Abundant experimental results suggest that adaptive weighting approach outperforms approaches simply weighting by sample proportion.

The rest of paper is organized as follows. We start with analyzing why weighting by sample proportion is not appropriate in some situations in Section 2. Then we introduce our adaptive weight optimization approach in Section 3. Section 4 presents experimental results. Conclusions are given in Section 5.

2 Motivation and Problem Statement

2.1 Motivation

Weighting by proportion (fixed weighting) could prevent a minority class from being ignored. However, as presented in previous research, the smaller number of instances in a training set, the less reliable corresponding classifier is[8]. In addition, from our observation and analysis, the more imbalanced the data are, the less reliable classifier with fixed weighting is. We illustrate this point from two perspectives: **proportion bias sensitivity** and **proportion bias asymmetry**.

Observation 1. *Class proportion bias is magnified in imbalanced data.*

For simplicity, we suppose a binary classification task with p positive instances and n negative instances. Imbalance rate is denoted as δ (i.e. $p = \delta n$). The ideal training set with total m instances should contain p' positive instances and n' negative instances where $p' : n' = \delta$. However, there may be d instances bias between random sampled training set and ideal training set. Consequently, imbalance rate we observed in training set is $(p' + d) : (n' - d) = \delta'$. We define **proportion bias sensitivity** to measure the effect of bias d .

$$s(\delta) = \frac{|\delta' - \delta|}{\delta} = \frac{\left| \frac{p'+d}{n'-d} - \delta \right|}{\delta} = \left| \frac{d\delta^2 + 2d\delta + d}{-d\delta^2 + m\delta - d\delta} \right| \quad (1)$$

It could be proved that $\forall \delta_1 > \delta_2, s(\delta_1) > s(\delta_2)$. This suggests that with the increasing of imbalance rate δ , proportion bias sensitivity is increasing monotonously.

Therefore, class proportion observed (δ') in training set may vary a lot from it of real set (δ) when class is extremely imbalanced.

Observation 2. *Original data set is more possible being less imbalanced than observed training set.*

Now we present that estimating class proportion from observed training set would be biased. Suppose we observed p positive instances and n negative instances in training set and sampling rate is $\frac{1}{r}$. The probabilities of real dataset having more than rp and less than rp (P_-) positive instances are asymmetry. This is referred to as **proportion bias asymmetry**. We set observed training set in four proportions (1:1, 3:1, 9:1, 49:1) and estimate probability of original data under the setting $p + n = 200, r = 5$. The result is presented in Figure 1. We could see that with the increasing of imbalance rate, probability curve becomes more asymmetric. Value of $P_+ : P_-$ is 1:1, 1:1.08, 1:1.19 and 1:1.56 respectively.

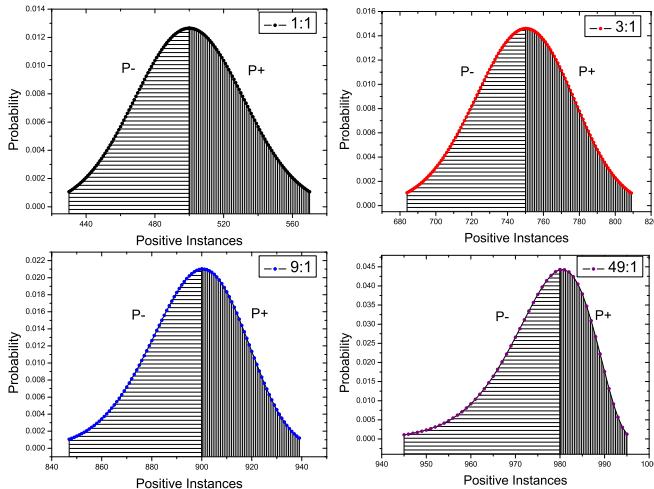


Fig. 1. Probability distribution of positive instances in real data set when the training set is observed at the imbalance rate of 1:1, 3:1, 9:1 and 49:1 respectively

2.2 Problem Statement

Considering a given training set $S = \{(\mathbf{x}_i, y_i)\}, i = 1, 2, \dots, n$ with n examples where $\mathbf{x}_i \in X$ is an instance in the m -dimensional feature space $X = \{f_1, f_2, \dots, f_m\}$ and $y_i \in Y = 1, 2, \dots, k$ is class label of instance \mathbf{x}_i . Task of classification is to find a classifier $g : X \rightarrow Y$ such that overall performance function $E(g)$ is maximal on testing set or real data set \widehat{S} . We suppose first p classes are majority classes and the others are minority classes. Then we could get $\forall_{i=1}^p \pi_i > 0 \gg \forall_{j=p+1}^n \pi_j$ where π_i denotes number of instances in classes y_i .

In fixed weighting approaches, it assigns a set of weights $\mathbf{w} = \{w_1, w_2, \dots, w_k\}$ to k classes by their proportions, $w_1 : w_2 : \dots : w_k = \frac{1}{\pi_1} : \frac{1}{\pi_2} : \dots : \frac{1}{\pi_k}$. Then it builds the best classifier $g_{\mathbf{w}}^* = \arg \max_{g_{\mathbf{w}}} E(g_{\mathbf{w}}, S)$ under weight distribution \mathbf{w} in

training set and assumes that performance of $g_{\mathbf{w}}^*$ in testing set $E(g_{\mathbf{w}}^*, \hat{S})$ is best as well. In previous, we demonstrated that $E(g_{\mathbf{w}}^*, \hat{S})$ may not be best because data distributions of S and \hat{S} are not exactly the same. In addition, differences of S and \hat{S} are magnified with the existence of data imbalance.

We propose an **adaptive weight optimization (AWO)** approach in this study. It tends to find the optimal weights \mathbf{w}^* and corresponding classifier $g_{\mathbf{w}^*}^*$ which could have best overall performance on all possible testing sets. Formally, we are searching for optimal weights \mathbf{w}^* from training set where

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} E(g_{\mathbf{w}}^*, S) \quad (2)$$

so that best classifier $g_{\mathbf{w}^*}^*$ under the optimal weights \mathbf{w}^* could outperform classifier with fixed weights on real data set \hat{S} .

3 Adaptive Weight Optimization Approach

3.1 Approach Overview

Here, we introduce overview of our approach. The first step is splitting training set S into two parts, namely, internal training set S_{tr} and the rest $S' = S - S_{tr}$. We further samples several internal testing sets S'_1, S'_2, \dots, S'_n from S' where $S'_i \subset S'$ to simulate distributions of possible testing sets. To make these internal testing sets more representative, we do both stratified sampling and random sampling. Next, we train a group of classifiers with initiate weight settings and use their performances on internal testing sets to guide weight optimization process. Finally, we could evaluate classifier under optimal weight on testing set.

3.2 Adaptive Weight Optimization

It is difficult to find the best weights directly. However, Evolutionary Algorithms provide multiple promising ways for optimization problem. Our method is inspired by the concept of Particle Swarm Optimization(PSO). It could also be designed based on other Evolutionary Algorithms such as Genetic Algorithm.

We encode a group of weights $\mathbf{w} = \{w_1, w_2, \dots, w_m\}$ as an particle. In initiation, we randomly produce several groups of weights as initial particles. Then we could build classifiers with weights in each particle by cost-sensitive methods. Next, we evaluate corresponding classifier on internal testing sets, and optimize each particle by overall performance. After plenty of iterations, particle with the best overall performance would be chosen as the optimal weight setting.

Detail optimization of each particle includes two steps in each iteration, namely, velocity updating and location updating. In our method, velocity is variation of weight and location is value of weight. Let $w_{i,j}(t)$ represent weight for class j in i^{th} particle and $v_{i,j}(t)$ for its variation at iteration t .

$$v_{i,j}(t+1) = g_w v_{i,j}(t) + c_1 R_1(pbest_{i,j} - w_{i,j}(t)) + c_2 R_2(gbest_j - w_{i,j}(t)) \quad (3)$$

$$w_{i,j}(t+1) = w_{i,j}(t) + v_{i,j}(t+1) \quad (4)$$

where g_w is inertia weight and c_1, c_2 are constrictions in PSO. We adopt g_w as a linear decreasing function and $c_1 = 2, c_2 = 2$ as suggested by standard PSO[1]. R_1 and R_2 are independent random numbers uniquely generated at every update for each individual dimension j . $pbest_{i,j}$ and $gbest_j$ are the previous best weight setting of particle i and the global best weight setting of all particles for class j respectively. They are determined by fitness function. For each particle i , we could get the best classifier $g_{\mathbf{w}_i}^*$ on training data. Then we evaluate each classifier on all internal testing sets and determine $pbest$ and $gbest$.

3.3 Fitness Function

Fitness function is critical in PSO because it governs updating of $pbest$ and $gbest$. Objective of imbalance classification is to build a classifier with relatively high precision and recall on all classes but not only on majority classes. We tried two assessment metrics here. They are Average-F1-Measure(Avg-F1) and extended G-Means(ex-GM). Suppose a k -classes problem, with definition of Precision, Recall and F1-Measure, two metrics are formulated as:

$$\begin{aligned} \text{Avg-F1} &= \sum_{i=1}^k \text{F1-Measure}(i)/k \\ \text{ex-GM} &= \left(\prod_{i=1}^k \text{Precision}(i) \cdot \text{Recall}(i) \right)^{\frac{1}{2k}} \end{aligned} \quad (5)$$

These two evaluation metrics all suggest that a classifier is considered good if it performs well on both majority classes and minority classes. We value two evaluation metrics equally in fitness function. Then we could compute fitness function with weight configuration \mathbf{w}_i on internal testing sets $\{S'_1, S'_2, \dots, S'_n\}$ by $fitness_i = \sum_{j=1}^n E(g_{\mathbf{w}_i}^*, S'_j)/n$. In addition, since we could observe proportion π_j of each testing set, we could estimate its probability $P(\pi_j)$ by *observation 2*. Fitness function could be formulated as:

$$fitness_i = \frac{\sum_{j=1}^n P(\pi_j) E(g_{\mathbf{w}_i}^*, S'_j)}{\sum_{j=1}^n P(\pi_j)} \quad (6)$$

In summary, procedures of adaptive weight optimization for multi-classification of imbalanced data is outlined in Algorithm 1 in pseudocode. Complexity of the algorithm is actually determined by speed of convergence and it is presented in Section 4.4.

4 Experiments

4.1 Experimental Setting

We set up abundant experiments to validate the usefulness of the proposed methodology. We take weighting by proportion as contrast. We couple these two weighting mechanisms with two cost-sensitive classification methods, namely, cost-sensitive decision tree(CSDT)[11] and cost-sensitive SVM(CSSVM) respectively. Due to limit of page length, we only report the results of CSDT. Results of CSSVM are very similar.

Algorithm 1. Adaptive weight optimization Algorithm

Input: Internal training and testing set S_{tr}, S'
Output: Classifier with optimal weight \mathbf{w}^*

```

1: //Sampling internal testing sets
2:  $n \leftarrow$  number of sampling sets;
3: for  $i = 1$  to  $n/2$  do
4:    $S'_i \leftarrow$ stratified sampling in each class of  $S'$ ;
5: end for
6: for  $i = n/2 + 1$  to  $n$  do
7:    $S'_i \leftarrow$ random sampling in  $S'$ ;
8: end for
9: //Adaptive weight optimization
10: for each particle do
11:    $\mathbf{w} \leftarrow$  initialize weight in range;
12: end for
13: for  $t = 1$  to  $maxIterations$  do
14:   for each particle do
15:     train classifier with  $\mathbf{w}_i$ ;
16:     compute fitness on all  $S'_j$ ;
17:     find  $pbest$  and  $gbest$ ;
18:      $v(t+1) \leftarrow$ Update( $v(t), w(t)$ );
19:      $w(t+1) \leftarrow$ Update( $w(t), v(t+1)$ );
20:   end for
21: end for

```

4.2 Comparison on Train-Test Benchmarks

We explore usefulness of adaptive weighting on given training set and testing set. Four data sets (*annealing*, *optdigits*, *pendigits*, *segment*) used in this section all come from UCI Machine Learning Repository. Class proportions of training set and testing set are near in *optdigits* and *pendigits* while they vary in some extent in *annealing* and *segment*. Table 2 shows the result on four data sets. We mark the better one in bold. In *optdigits* and *pendigits* data set, performances of fixed weighting and adaptive weighting are similar. It is exactly the same on *pendigits*. In other two data sets, adaptive weighting shows its advantages. It confirms our claims that adaptive weighting could perform well when distribution of testing set is different with training set and it is effective when data is imbalanced. As we analyzed in previous section, sampling proportion in training set may vary a lot with testing set. Fixed weighting could only ensure best performance in training set. But it is not effective in testing set due to different sample proportion. We also employ SMOTE and under-sampling method in comparison. Cost-sensitive with adaptive weighting also outperforms data level methods. It implies that algorithm level method is promising if cost is chosen appropriately.

4.3 Comparison on Imbalance Rate

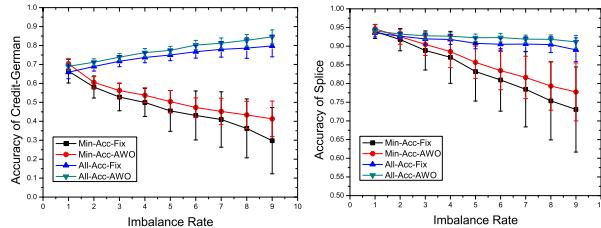
We validate the claim that adaptive weighting is more effective when data is extremely imbalanced here. We manually re-sample original data set at different imbalance rate. We set imbalance rate of majority classes to minority classes from 1:1 to 9:1 by stratified sampling. Due to the limitation of page length, we only report results on two data sets (*Credit-German* and *Splice*) for illustration.

The results are shown in Figure 3. Left figure is result on a binary classification task and right one is a multi-classification task. We compared overall

Table 1. Performance comparisons on train-test data sets

		Avg-F1	ex-GM	Avg-Pre	Avg-Re	ACC
optdigits	fix	0.955	0.931	0.957	0.955	0.955
	adaptive	0.962	0.941	0.965	0.961	0.962
pendigits	fix	0.981	0.947	0.983	0.980	0.980
	adaptive	0.981	0.947	0.983	0.980	0.980
Annealing	SMOTE	0.877	0.831	0.828	0.939	0.878
	Under sampling	0.868	0.831	0.825	0.938	0.877
	fixed	0.860	0.824	0.815	0.935	0.893
	adaptive	0.887	0.846	0.842	0.944	0.917
segment	SMOTE	0.972	0.972	0.971	0.969	0.968
	Under sampling	0.968	0.968	0.969	0.967	0.965
	fix	0.962	0.962	0.964	0.964	0.962
	adaptive	0.973	0.973	0.978	0.972	0.974

accuracy and accuracy of minority classes under different imbalance rate. Adaptive weighting could derive better performance on all imbalance rates. Moreover, advantage of adaptive weighting over fixed weighting becomes significant when imbalance rate increases. From standard deviation, it is obvious that classifier with adaptive weighting is more consistent. In every run, we randomly split resampled data set into training set and testing set. Accuracies of fixed weighting are highly dependent on the distributions of training set and testing set. Result of fixed weighting varies a lot in each run. Adaptive weighting which is suitable for most of possible testing set distributions so that its performances are more stable. The results again demonstrate the usefulness of adaptive weighting in reaching high overall accuracy as well as accuracy of minority classes.

**Fig. 2.** Overall accuracy and minority accuracy comparisons of fixed weighting and adaptive weighting on different imbalance rate

4.4 Speed of Convergence

Speed of convergence is very important in Evolutionary Algorithms. In most of our experiments, the weight distribution converges in less than 10 iterations. Maximal number of iterations is 16. Since slight variation of weight would not influence performance a lot, it is very fast in speed of convergence. In addition, we find that converged weight distribution is different with fixed weighting, especially when imbalance rate is high.

5 Conclusion

In this study, we have discussed classification of imbalanced data. We analyzed drawbacks of traditional fixed weighting by class proportion. To overcome the problem, we proposed an adaptive weighting approach which could optimize weights according to data distributions. In detail, we formulated it as an optimization problem and employed Particle Swarm Optimization to search for optimal weights. Abundant experiments are conducted to validate proposed approach. Improvements have been observed on a range of benchmark classification tasks. Advantage of adaptive weighting would become obvious with the increasing of imbalance rate.

References

1. Bratton, D., Kennedy, J.: Defining a standard for particle swarm optimization. In: IEEE Swarm Intelligence Symposium, SIS 2007, pp. 120–127. IEEE (2007)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. In: Journal of Artificial Intelligence Research, pp. 321–357 (2002)
3. Fernández, A., García, S., Herrera, F.: Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In: Hybrid Artificial Intelligent Systems, pp. 1–10 (2011)
4. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
5. Jan, T.K., Wang, D.W., Lin, C.H., Lin, H.T.: A simple methodology for soft cost-sensitive classification. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 141–149. ACM (2012)
6. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(2), 539–550 (2009)
7. Masnadi-Shirazi, H., Vasconcelos, N.: Risk minimization, probability elicitation, and cost-sensitive svms. In: Proceedings of the International Conference on Machine Learning, pp. 204–213 (2010)
8. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. The MIT Press (2009)
9. Rosenberg, A.: Classifying skewed data: Importance weighting to optimize average recall (2012)
10. Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L.: Cost-sensitive learning methods for imbalanced data. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2010)
11. Ting, K.M.: An instance-weighting method to induce cost-sensitive trees. IEEE Transactions on Knowledge and Data Engineering 14(3), 659–665 (2002)
12. Weiss, G.M.: Mining with rarity: a unifying framework. Sigkdd Explorations 6(1), 7–19 (2004)
13. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE International Conference on Data Mining, ICDM 2003, pp. 435–442. IEEE (2003)

Feature Selection via Sparse Regression for Classification of Functional Brain Networks

Yilun Wang¹, Guorong Wu², Zhiliang Long², Jingwei Sheng², Jiang Zhang³,
and Huafu Chen^{2,*}

¹ School of Mathematical Sciences,
University of Electronic Science and Technology of China, Sichuan, China
yilun.wang@gmail.com

² Key laboratory for Neuroinformation of Ministry of Education,
School of Life Science and Technology,
University of Electronic Science and Technology of China, Chengdu, P.R. China
chenhf@uestc.edu.cn

³ Information Research Institute, Southwest Jiaotong University,
Chengdu 610031, China

Abstract. Despite the ongoing progress to chart the differences between the healthy controls and patients at the group level, the pattern classification of functional brain networks across individuals is still a challenging task. The difficulties include the very high dimensional feature space and very small sample size, as well as the probably high noise level. In this paper, we apply the stable sparse regression to pick the very few most discriminant features (edges) for the following classification. We considered different noise to signal ratios and sparsity controlling parameters and numerical experiments based on simulated data demonstrate the much better classification performance via the feature selection based on the sparse regression.

Keywords: sparse regression, feature selection, stability selection, classification.

1 Introduction

The human brain is among the most complex network systems in the world, considering that it comprises about one hundred billion neurons, with thousands of trillions of connections between them. The anatomical and physiological studies in past few decades provided a significant body of evidence for the important role of structural connectivity in shaping physiological responses. Meanwhile, functional connections that describe statistical dependencies are derived from observations of neural time series, reflecting functional segregation and localization of function in neuroscience. That is to say, the human brain can be considered as a large-scale network, with nodes being distinct brain regions and edges representing functional or structural connectivity among them [1, 2].

* Corresponding author.

In this paper, we are focusing on pattern recognition and classification based on the brain network. Most existing research on brain networks simply focuses on describing group differences between subject classes (knowing the label of each subject) and cannot classify or predict the brain behavior across individuals, due to the relatively small number of subjects, very high dimensional feature space (consisting of the network edges) and probably high level noise, leading to the over-fitting training data and curse of dimensionality problem. Therefore, in this paper, we are focusing on selecting a small number of most discriminative features, to significantly reduce the dimension of the feature space and correspondingly enhance the classification performance. It is quite reasonable to perform feature (edge) selection, because usually only a small proportion of the pathways in the brain might be responsible for the dysfunction or certain task of the brain network.

Recently, sparse modeling, as a rapidly developing area at the intersection of statistics, machine-learning and signal processing [3, 4], can find out a small number of the most relevant variables in a high-dimensional feature space and therefore is most appealing for practical feature selection. It has been applied to many problems, including the voxel selection to localize brain activation patterns corresponding to different stimulus classes or brain states [5–7].

In this paper, we study the application of sparse regression to the feature (edge) selection in brain network with an aim to identify a small proportion of the discriminative functional pathways and brain regions. While there have existed some related work [8, 9], the deep study of sparse regression for feature selection on the brain network is still very limited. For example, the effects of the different signal and noise ratios and the discussion of the different sparsity levels have not been considered. In this study, we will deepen the existing study in the above two aspects. Notice that while our study is not limited to a specific type of network, we are mainly focusing on the network based on the functional MRI, and will mainly use the simulated data, since our main goal is to evaluate the methodology of the stable sparse regression.

2 Methods

2.1 Subjects

In total, the fMRI data of 100 subjects is generated and equally divided into 2 groups, i.e. Group 1 and Group 2, respectively. Group 1 and Group 2 differs with each other mainly in terms of the strength of functional connectivity between certain regions and we will explain it in more details later. As we know, the simulation data is usually designed to facilitate the deep understanding and testing of a variety of analytic or computational methods before they can be applied to the real data.

We adopt a data generation model that is consistent with the spatiotemporal separability assumptions of independent component analysis (ICA), that is, data can be expressed as the product of time courses (TCs) and spatial maps (SMs). For each subject, the spatial map is the same, because we are considering

the connectivity between same brain regions. Here we are considered 256 brain regions, and the 256 regions are further divided into two categories, i.e. 87 active regions and 169 non-active regions. They differ in terms of the definition of the corresponding time courses. For each region, we define its average time course with T time points in length. Its construction is under the assumptions that component activations result from underlying neural events as well as noise. In this simulation, each time course has $T = 160$ time points. For the active regions, the time course is divided into several blocks. The signal value of each task block is set to be a positive value between 0.9 to 1, while the signal value for the resting blocks is set to be 0. As the non-active regions, the time course is defined using random time series, defined by normal distribution with mean being 1 and the standard deviation being 1. In order to test the robustness of the classifiers, we add different levels of Gaussian noise.

The functional connectivity is established by calculating the covariances of the time courses between different regions. We have 256 regions and correspondingly the generated network of each subjects has $256 \times 255/2 = 32640$ edges. Groups 1 differs with Groups 2 in terms of the functional connectivity strength of 6 edges.

2.2 Sparse Logistic Regression

In this paper, we are considering the functional connectivity of each subject and correspondingly the feature vector consists of all the edges. The number of edges is typically very big in practice and it is 32640 in our simulated data. That is to say, each subject is defined in a 32640-dimensional feature space. The high dimension of the feature space often bring difficulty for the following classifiers no matter in terms of computational cost or classification accuracy. Therefore we want to perform dimensional reduction by selecting the small number of most discriminative edges, which are used to form a much lower dimensional new feature space.

We use sparse Sparse Logistic Regression based supervised learning to perform the feature selection. The basic formulation of the sparse logistic regression is

$$\min_x \|Ax - y\|_2^2 + \lambda \|x\|_1, \quad (1)$$

where A is the training data matrix, where each row is a 32640 dimensional vector representing one subject, and each column represents a feature; x is the unknown regression coefficients; y is the known label vector for the training data, with 1 representing the subject in Groups 1 and -1 for the Groups 2. The ℓ_1 norm $\|x\|_1$ is the sparsity regularization term, and λ is the regularization parameter that controls the degree of sparsity. A larger λ leads to x with more zeros, i.e. a more sparse coefficient vector. There have existed many different solvers for the above sparse logistic regression model and in the paper we use the solver "SLEP: A Sparse Learning Package" developed by the Arizona State University [10]. By solving the sparse regression problem, we obtain the regression coefficients x whose absolute value indicates the contribution of the corresponding edges to discriminating these two groups.

Randomization for Stability Selection. Due to the presence of the correlation in the training data matrix A , we adopt the randomized sparse regression called stability selection [13] to improve the performance of the sparse feature selection method. The randomized sparse regression is to repeat solving many similar sparse regression problems as (1). Each problem is generated by randomly perturbing the data matrix A via taking only a fraction of the training samples and randomly scaling each feature, in our case each edge. By counting how often each edge is selected across the repetitions, each edge can be assigned a score. Higher scores denote variables likely to belong to the set of the true discriminative edges.

2.3 Support Vector Machine

There have existed many classifiers and in this paper we take the widely used Support Vector Machine (SVM) as an example to demonstrate that our feature selection can help improve the classification accuracy of SVM [11].

Support Vector Machine (SVM) is a specific type of supervised machine learning method that aims to classify data points by maximising the margin between classes in a high-dimensional space and it adopts the ℓ_2 -norm regularization to avoid over-fitting. This optimization problem belongs to quadratic programming and can be efficiently solved by many specific solvers such as sequential minimal optimization. Like most of classification methods, SVM involves the training stage and the testing stage. The goal of SVM is to produce a model based on the training data to predict the target values of the testing data. The traditional SVM performs linear classification, but non-linear classification can also be performed by incorporating the so called the kernel trick, which implicitly maps the inputs into high-dimensional feature spaces. Due to its outstanding practical performance and solid theory guarantee based on the statistical learning, SVM becomes one of the most popular classifiers and therefore in this paper we use it to demonstrate the performance of our feature selection scheme based on sparse optimization.

Consider a training data set $D = \{(x_i, y_i), i = 1, \dots, n\}$ where $x_i \in R^d$ are data points and y_i are labels. The problem of learning a linear classifier, $y = \text{sign}(\omega^T x + b)$, where $y = \{1, -1\}$ or a linear function $y = w^T + b$ where y is a scalar can be understood as estimating $\{\omega, b\}$ from D . Over the years Support Vector Machines(SVMs) have emerged as powerful tools for estimating such functions.

To develop notation we briefly discuss the problem of training linear classifiers. The SVM formulation for linearly separable datasets is given by

3 Numerical Results

In this paper, the classification accuracy is measured by the commonly used quantities, such as generalization rate (GR), sensitivity (SS) and specificity (SC). Here The proportion of all subjects that were correctly predicted is evaluated by

the GR; SS is defined as the proportion of correctly predicted Group 1 subjects, while SC represents the proportion of correctly predicted Group 2 subjects. Their formulations are shown below:

$$GR = (TP + TN) / (TP + FN + TN + FP)$$

$$SS = TP / (TP + FN)$$

$$SC = TN / (TN + FP),$$

where TP is the number of the Group 1 subjects correctly predicted; FN is the number of the Group 1 subjects classified as in Group 2; TN is the number of the Group 2 correctly predicted; FP is the number of the Group 2 subjects classified as in Group 1 [12]. In this paper, we are using the leave-one-out cross-validation, which use a single subject as the test data and all the remaining as the training data. Each of the 100 subjects is chosen as the test data in turn without missing or repetition, and finally we calculate the value of TP, FN TN, FP, where $TP + FN + TN + FP = 100$.

We use the Support Vector Machine (SVM) as the classifier, which is provided by the toolbox of MATLAB 2012b and the default parameters are used. We first do not perform the feature selection and use all the 32640 edges as the input of SVM, run SVM and record the classification results. Then we first perform edge selection by sparse regression, and then only use the information of the very few number of selected edges as the input of SVM, run SVM, and record the classification result. We compare these two classification methods and demonstrate the significant role of selection of discriminative edges for the improvement of classification performance, in cases of adding different levels of noises.

The performance of the classifier was estimated using leave-1-out validation test with an 100 times repetition. We carried out a simulation study to assess (i) the classification performance under various Noise to Signal Ratio (NSR, for short), which is the reciprocal of the Signal to Noise Ratio (SNR, for short); (ii) the effect of different choices of the penalty parameter λ on our ability to detect the most discriminative interaction. The classification results are summarized in Table 1, Table 2, Table 3 and Table 4. "W/" represent the classification results of SVM together with the feature selection via sparse regression while the "W/O" represent the classification results of SVM without the feature selection via sparse regression. As for the regularization parameter λ , Table 1 and Table 2 are for $\lambda = 1.5$ while Table 3 and Table 4 are for $\lambda = 0.015$. As for the number of selected features (edges, here), Table 1 and Table 3 are for 6 selected most discriminative edges while Table 2 and Table 4 are for 12 selected edges. Here we note that in practice, ones might not know the exact number of the most discriminative edges as we did for simulation data. However, in many simulations ones have a rough estimation based on their experiences and performing feature (edge) selection on this number is still helpful. We will see that selecting 12 edges instead of 6 edges still brings great improvement of classification accuracy.

From the results of Table 1, 2, 3 and 4, we have several preliminary observations. 1) The performance of feature selection in terms of classification is not

Table 1. Classification results where $\lambda = 1.5$ and number of selected features is 6

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.99	0.82	0.99	0.84	1.00	0.80
0.8	0.96	0.73	0.98	0.82	0.94	0.64
1.0	0.95	0.76	0.92	0.84	0.98	0.68
1.2	0.88	0.67	0.90	0.72	0.86	0.62
1.6	0.75	0.53	0.72	0.52	0.78	0.54

Table 2. Classification results where $\lambda = 1.5$ and number of selected features is 12

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	1.00	0.82	1.00	0.84	1.00	0.80
0.8	0.96	0.73	0.98	0.82	0.94	0.64
1.0	0.92	0.76	0.88	0.84	0.96	0.68
1.2	0.78	0.67	0.74	0.72	0.82	0.62
1.6	0.69	0.53	0.72	0.52	0.66	0.54

Table 3. Classification results where $\lambda = 0.015$ and number of selected features is 6

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.99	0.82	0.98	0.84	1.00	0.80
0.8	0.97	0.73	1.00	0.82	0.94	0.64
1.0	0.89	0.76	0.90	0.84	0.88	0.68
1.2	0.88	0.67	0.90	0.72	0.86	0.62
1.6	0.81	0.53	0.82	0.52	0.80	0.54

Table 4. Classification results where $\lambda = 0.015$ and number of selected features is 12

NSR	GR		SS		SC	
	W/	W/O	W/	W/O	W/	W/O
0.2	1.00	0.89	1.00	0.99	1.00	0.80
0.4	1.00	0.87	1.00	0.94	1.00	0.80
0.6	0.96	0.82	0.96	0.84	0.96	0.80
0.8	0.97	0.73	0.96	0.82	0.98	0.64
1.0	0.94	0.76	0.94	0.84	0.94	0.68
1.2	0.83	0.67	0.82	0.72	0.84	0.62
1.6	0.76	0.53	0.80	0.52	0.72	0.54

strongly dependent on the choice of λ and therefore the sparse regression is reliable in practice, though the high noise level might prefer smaller λ while the low noise level might prefer larger one. 2) As the noise-to-signal ratio increases, the recognition performance deteriorate as expected, but the feature selection via sparse regression always brings significant better recognition accuracy. 3) The performance of feature selection in terms of classification is not strongly dependent on the prescribed number of selected features, if the adopted number is not far away from the true number of significantly discriminative features.

Acknowledgement. This work was supported by the Natural Science Foundation of China, Grant Nos. 61035006, 61125304, 11201054, 61273361 and by the Specialized Research Fund for the Doctoral Program of Higher Education of China 20120185110028, and by the Fundamental Research Funds for the Central Universities ZYGX2012J118, and by the Key Technology R&D Program of Sichuan Province 2012SZ0159.

References

1. Friston, K.J.: Functional and Effective Connectivity: A Review. *Brain Connectivity* 1(1), 13–36 (2011), doi:10.1089/brain.2011.0008.
2. Sporns, O.: The Human Connectome: Origins and Challenges (to appear 2013)
3. 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* 227(2), 210 (2009)
4. Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences* 1(3), 248–272 (2008)
5. Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A.: Multivoxel Pattern Analysis for fMRI Data: A Review. *Computational and Mathematical Methods in Medicine*, Article ID 961257, 14 (2012)
6. Li, Y., Namburi, P., Yu, Z., Guan, C., Feng, J., Gu, Z.: Voxel Selection in fMRI Data Analysis Based on Sparse Representation. *IEEE Transaction on Biomedical Engineering* 56(10) (2009)
7. Li, Y., Long, J., He, L., Lu, H., Gu, Z., et al.: A Sparse Representation-Based Algorithm for Pattern Localization in Brain Imaging Data Analysis. *PLoS ONE* 7(12) (2012)
8. Zhang, J., Cheng, W., Wang, Z., Zhang, Z., Lu, W., Lu, G., Feng, J.: Pattern Classification of Large-Scale Functional Brain Networks: Identification of Informative Neuroimaging Markers for Epileps. *PLoS ONES* 7(5) (2012)
9. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage* 44, 112 (2009)
10. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections, Arizona State University (2009), <http://www.public.asu.edu/~jye02/Software/SLEP>
11. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3), 27:1–27:27 (2011), <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
12. Su, L., Wang, L., Chen, F., Shen, H., Li, B., et al.: Sparse Representation of Brain Aging: Extracting Covariance Patterns from Structural MRI. *PLoS ONE* 7(5) e36147 (2012), doi:10.1371/journal.pone.0036147
13. Meinshausen, N., Bühlmann, P.: Stability selection. *J. Roy. Statistical Society B* 72, 417 (2010)

Efficient Euclidean Local-Structural Based Sparse Coding for Robust Visual Tracking*

Ji Zhang^{1,2}, Hong-Yuan Wang^{1,2,✉}, and Fu-Hua Chen³

¹ School of Information Science and Engineering, ChangZhou University,
ChangZhou 213164, China

² ChangZhou Key Laboratory for Process Perception and Interconnected Technology,
ChangZhou 213164, China

³ Department of Natural Science and Mathematics, West Liberty University,
West Virginia 26074, United States
 {zhangji,hywang}@cczu.edu.cn, fuhua@wlu.edu

Abstract. Recently, sparse coding based visual tracking(ℓ_1 -tracker) has been obtained increasing attention and many relative algorithms are proposed. In the framework of these algorithms, each candidate region is sparsely represented by a set of target templates. However, almost all these algorithms ignore the structural information among candidate regions. Lu proposes NLSSC-tracker based on non-local self-similarity constraint, but it falls into the high computational costs like that in ℓ_1 -tracker. In this paper, we consider an Euclidean local-structural constraint as smooth operator, and propose a novel ELSSC-tracker. Our optimization procedure proposed in this paper can be transformed to be a small-scale ℓ_1 -problem skillfully. Extensive experimental results have demonstrated the effectiveness and efficiency of our algorithm.

Keywords: Euclidean Local-Structure Constraint, ℓ_1 -tracker, Sparse Coding, Target Tracking.

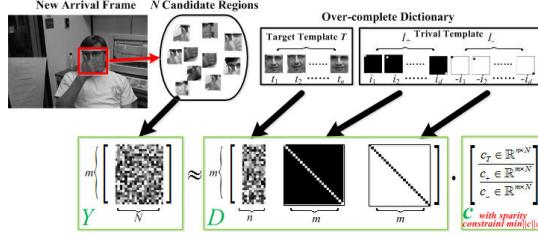
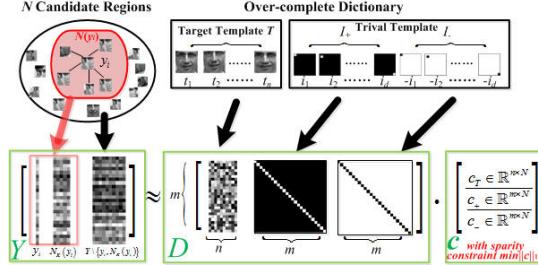
1 Introduction

Recently, visual tracking has been widely used in security surveillance, navigation, human-computer interaction and so on[1][2]. Targets for tracking change dynamically and uncertainly, because of occlusion, noisy, varying illumination and so on. In the last twenty years, many tracking algorithms have been proposed, which can be divided into two categories: generative tracking and discriminant tracking[1], and our algorithm in this paper belongs to the former.

Recently, Mei proposes ℓ_1 -tracker for generative trackin[4], and occlusion, corruption and other challenging issues presented in tracking are addressed seamlessly. However, in ℓ_1 -tracker, the computational costs are too expensive to achieve efficient tracking, and the local structures of similar regions are ignored;

* The National Natural Foundation of China under Grant No. 61070121,60973094.

✉ Corresponding author.

**Fig. 1.** Original ℓ_1 -tracker Algorithm**Fig. 2.** Lu's NLSSSC-tracker Algorithm

Xu considers two desired properties of ℓ_1 -regularization: sparsity and algorithmic stability[5], which are fundamentally at odds with each other; Lu proposes NLSSSC-tracker which takes the structural information of the target templates as a smooth operator[6]. However, Lu's tracker is too slow to offer real-time tracking. In this paper, motivated by the robustness of ℓ_1 -tracker and the stability of NLSSSC-tracker, we propose a novel ELSSC-tracker, which performs robustly and efficiently. The rest of this paper is organized as follows: sparse coding and ℓ_1 -tracker are introduced in section 2; In section 3, we analysis the disadvantages of ℓ_1 - and NLSSC-tracker, and propose our ELSSC-tracker; Experimental results are reported in section 4 and the conclusion and future work are summarized finally in Section 5.

2 Relative Works

2.1 Sparse Coding and ℓ_1 -tracker

Sparse coding(SC for short, also named as sparse sensing or compressive sensing), is an attractive signal reconstruction method, proposed by Candès[3]. The main task of SC is to reconstruct a signal $y \in \mathbb{R}^{m \times 1}$ with over-completing dictionary $D \in \mathbb{R}^{m \times n}$ with sparse coefficient vector $c \in \mathbb{R}^{n \times 1}$. The formulation of ℓ_1 -optimization based on SC can be written as following optimization problem:

$$\min_c \|y - Dc\|_F^2 + \alpha \|c\|_1 \quad (1)$$

where $\|c\|_1$ is the summary of entity elements in c . Many methods have been proposed for solving it, including basic pursuit(BP), orthogonal matching pursuit(OMP), least absolute shrinkage and selection operator(LASSO), etc[3].

Based on SC, Mei presents ℓ_1 -tracker and the procedure of which is shown in Fig.1. Suppose that, target has been located in the last frame, ℓ_1 -tracker is initialized with it in the new arrival frame and N candidate regions are generated around it, as shown in Fig.1. With n templates learned from previous tracking and $2m$ trivial templates including m positive ones and m negative ones(m is the dimension of $1D$ stretched image), Eq.(1) can be solved as the second row of Fig.1. And with positive and negative trivial templates, Mei adds non-negative constraint $c \geq 0$ in Eq.(1). Reconstruction errors of all candidate regions with sparse coding coefficients can be used to determine the weights for each candidate, and object in the new arrival frame can be located with the sum of weighted candidates. The updating strategies of dictionaries can be seen in[4].

2.2 Non-Local Self-Similarity Based Tracking(NLSSSC-tracker)

Recently, Xu presents that, sparsity and stability are in a dilemma in sparse regularized algorithms[5]. There are lots of similar regions in all N candidates, Lu proposes the non-local self-similarity(NLSS) constraint as following

$$\sum_{i=1}^n \|c_i - \sum_j w_{ji} c_j\|^2 = \|C - CW\|_F^2 \quad (2)$$

where c_i, c_j are sparse coefficients for candidate regions y_i, y_j . Given a set of m -dimensional candidates $Y = [y_1, \dots, y_N]$, y_j is selected if it is within the first K closest points of y_i (denoted as $N_K(y_i)$). The weight w_{ij} can be denoted as $w_{ji} = \frac{1}{s_i} e^{-\frac{\|N(y_i) - N(y_j)\|^2}{h}}$, where h is a parameter enforcing the similarity and s_i is the normalization factor. Lu's NLSSSC-tracker can be formulated as

$$\min_C \|Y - DC\|_F^2 + \alpha \|C\|_1 + \beta \|C - CW\|_1 \quad (3)$$

Taking Mei's ℓ_1 -tracker in Eq.(1) as initial coefficients C_0 , Eq.(3) can be solved iteratively[6]. The high computational costs of original ℓ_1 -tracker and the iterative procedure for maintaining the neighborhood constraints of sparse coefficients make Lu's NLSSSC-tracker hard to achieve real-timing tracking.

3 Euclidean Local-Structural Based Sparse Coding for Tracking(ELSSC-tracker)

Confronting with the high time-consuming of ℓ_1 - and NLSSSC-tracker, we propose an efficient ELSSC-tracker with Euclidean local-structural based sparse coding in this section, to mention the local structures of candidate regions. The main features of our ELSSC-tracker are stable, sparse, efficient and robust.

3.1 Euclidean Local-Structural Constraint Based Sparse Coding

It is clear to see in Eq.3, NLSSSC is a double ℓ_1 -problem. But it is well known that, ℓ_2 -optimization is much easy to solve than ℓ_1 -optimization, thus we use ℓ_2 -norm to measure the relationships of sparse coefficient vectors, and remain ℓ_1 -constraint to maintain the sparsity of optimization as following

$$\min_C \|Y - DC\|_F^2 + \alpha \|C\|_1 + \beta \|C - CW\|_F^2 \quad (4)$$

Eq.(4) can be solved iteratively. For a single candidate region y_i , in the m -th iteration of optimization, Eq.(4) can be written as

$$\min_{c_i^{(m)}} f(c_i^{(m)}) = \min_{c_i^{(m)}} \|y_i - Dc_i^{(m)}\|_2^2 + \alpha \|c_i^{(m)}\|_1 + \beta \|c_i^{(m)} - \theta_i^{(m-1)}\|_2^2 \quad (5)$$

where $\theta_i^{(m-1)} = \sum_j w_{ji} c_j^{(m-1)}$. As in the m -th iteration of c_i , $c_{j,j \neq i}$ are fixed, we can regard $\theta_i^{(m-1)}$ as a constant. In order to solve Eq.(5), we introduce the following surrogate function like that in [6]

$$\psi(c_i, c_0) = \frac{\lambda}{2} \|c_i^{(m)} - c_0\|_2^2 - \frac{1}{2} \|Dc_i^{(m)} - Dc_0\|_2^2 \quad (6)$$

where λ is chosen to make $\psi(c_i, c_0)$ convex. As Daubechies's proof in [7], when $\lambda I - D^T D$ is a strictly positive-definited matrix, $\psi(c_i, c_0)$ is strictly convex for any choice of c_0 . Hence, in our experiments, we pick a constant λ so that $\|D^T D\| < \lambda$, as shown in Algorithm.1. In the following derivation, we set $\lambda = \gamma - 2\beta$ for convenience. For a single candidate y_i , the surrogate objective function is

$$\begin{aligned} f(c_i^{(m)}) &= \frac{1}{2} \|y_i - Dc_i^{(m)}\|_2^2 + \alpha \|c_i^{(m)}\|_1 + \beta \|c_i^{(m)} - \theta_i^{(m-1)}\|_2^2 \\ &\quad + \frac{\gamma - 2\beta}{2} \|c_i^{(m)} - c_0\|_2^2 - \frac{1}{2} \|Dc_i^{(m)} - Dc_0\|_2^2 \\ &= \frac{1}{2} \|y_i\|_2^2 - \langle y_i, Ds_i^{(m)} \rangle + \alpha \|c_i^{(m)}\|_1 - 2\beta \langle c_i^{(m)}, \theta_i^{(m-1)} \rangle \\ &\quad + \frac{\gamma}{2} \|c_i^{(m)}\|_2^2 - (\gamma - 2\beta) \langle c_i^{(m)}, c_0 \rangle + \frac{\gamma - 2\beta}{2} \|c_0\|_2^2 \\ &\quad + \langle Dc_i^{(m)}, Ds_0 \rangle - \frac{1}{2} \|Dc_0\|_2^2 \\ &= \frac{\gamma}{2} \|c_i^{(m)} - v_i^{(m)}\|_2^2 + \alpha \|c_i^{(m)}\|_1 + R \end{aligned} \quad (7)$$

where, $v_i^{(m)} = \frac{1}{\gamma} [D^T y_i + 2\beta \theta_i^{(m-1)} + (\gamma - 2\beta) c_0 - D^T D c_0]$ and $R = \frac{1}{2} \|y_i\|_2^2 + \frac{\gamma - 2\beta}{2} \|c_0\|_2^2 - \frac{1}{2} \|Dc_0\|_2^2 - \frac{\gamma}{2} \|v_i^{(m)}\|_2^2$ is a constant in m -th iteration. We can simplify Eq.(7) into

$$f(c_i^{(m)}) = \frac{\gamma}{2} \|c_i^{(m)} - v_i^{(m)}\|_2^2 + \alpha \|c_i^{(m)}\|_1 \quad (8)$$

We can consider Eq.(8) as a ℓ_1 -optimization. Firstly, we decompose over-completing dictionary D in Eq.(4) with SVD as $D = U \Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$,

$\Sigma \in \mathbb{R}^{m \times n}$, $V \in \mathbb{R}^{n \times n}$. It is clear to see that, V is a orthogonal matrix, and Eq.(8) can be rewritten as

$$f(c_i^{(m)}) = \frac{\gamma}{2} \|x_i^{(m)} - Vc_i^{(m)}\|_2^2 + \alpha \|c_i^{(m)}\|_1 \quad (9)$$

where, $x_i^{(m)} = Vv_i^{(m)}$. With this, we can transform the optimization problem with ELS constraint in Eq.(5) into a pure ℓ_1 -minimization problem in Eq.(9), and the minimization can be seen as to represent the signal $x_i^{(m)}$ with sparse coefficients $c_i^{(m)}$ under the dictionary V , it is completely different from the method used by Lu in [6], named Euclidean Local-structural based Sparse Coding(ELSSC for short). Experiments show that, our ELSSC converges with 1-2 iterations at most, while Lu's method with more than 3 under the same settings.

3.2 Our Original and Improved ELSSC-tracker

In this section, based on the above optimization procedure of ELSSC, we propose our ELSSC-tracker under the framework of original ℓ_1 -tracker[4]. The main difference between ℓ_1 -tracker, NLSSSC-tracker and ELSSC-tracker is that, we need to iteratively solve 2-3 times large scale ℓ_1 -optimization of Eq.(9) for each candidates in our algorithm, while once in ℓ_1 -tracker and in NLSSSC-tracker, which means our tracker is much slower than other two ones. When the target template is sized 40×40 and the over-completing dictionary D is sized 1600×3210 (parameters in our experiments), the orthogonal matrix $V \in \mathbb{R}^{3210 \times 3210}$.

In order to reduce the scale of optimization, we need to reduce the size of V (the new dictionary in Eq.(9)) primarily. Due to the property of SVD and the structure of D (the original dictionary in Eq.(5), the last 3200 columns are constructed by a 1600-ordered positive and a 1600-ordered negative identity matrix), we can get the conclusion that, $D = U\Sigma V^T = U\Sigma_{1:n}V_{1:n}^T$, where $\Sigma_{1:n}$ and $V_{1:n}$ denote the first n columns of Σ and V respectively. Thus, each iteration of every candidate regions in our original ELSSC-tracker can be reduce from the huge scale ℓ_1 -optimization to a much smaller one($V \in \mathbb{R}^{n \times n}$, where $n=10$ in our experiments). And in order to cover the occlusions in tracking, we take the similar trivial templates(a n -dimensional identity matrix and a n -dimensional negative identity matrix) to strength the dictionary V in Eq.(9), but our new dictionary $V' \in \mathbb{R}^{10 \times 30}$ is still much smaller than $V \in \mathbb{R}^{3210 \times 3210}$ in original ELSSC-tracker and $D \in \mathbb{R}^{1600 \times 3210}$ in original ℓ_1 -tracker.

4 Experiments

4.1 Experiment Setting

In order to evaluate our tracker, we conduct experiments on Surfer, Dudek, Faceocc2 and Animal with 375, 1145, 819 and 71 frames respectively. These sequences cover almost all challenges in tracking, including occlusion, motion blur, rotation, scale variation and complex background, etc. For comparison, we use

four state-of-the-art algorithms with the same initial positions and representation of targets for tracking, including incremental learning based tracker(IVT, a common discriminant tracker)[8], covariance based tracker(CovTracking, a generative tracker on Lie-group)[9], sparse coding based original ℓ_1 -tracker(a generative tracking methods)[4] and original NLSSC-tracker[6]. Experiments are running on computer with 2.67GHz CPU and 2GB memory.

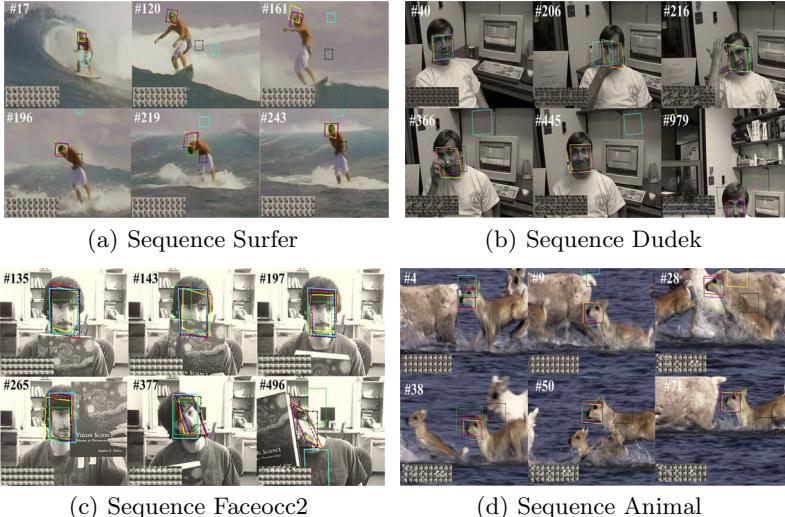


Fig. 3. Some Tracking Results

The main parameters used in our relative experiments are setting as following: the number of particles $N=200$, the number of template regions $n=10$, and the size of candidates and targets are reshaped to 40×40 ; ℓ_1 -optimizations used in ℓ_1 -tracker and our ELSSC-tracker are solved with mexLASSO with $\alpha=0.01$; β in Eq.(7) is 0.5, and γ is setting dynamically when dictionary D in Eq.(7) is updated after getting the target region under the constraint that $\|D^T D\| \leq \lambda = \gamma - 2\beta$; c_0 in Eq.(7) is checkless, and we set it as all 0-vector for convenience.

4.2 Experimental Results for Visual Target Tracking

We evaluate the above-mentioned algorithms using the center location errors, average success rates and average frames per second, and the results are shown in Fig.3, Fig.4 and Tab.1. Templates of NLSSC-, original ELSSC- and improved ELSSC-trackers are shown with three rows in the left-bottom of Fig.3.

For occlusion, five algorithms except IVT work steadily roughly, especially at #206,#366 of Dudek sequence in Fig.3(b)(head for tracking is covered by hand and glasses) and #143,#265,#496 of Faceocc2 sequence in Fig.3(c)(head for tracking is covered by book), and after the target recovers from occlusion, these trackers can seek it quickly. IVT works poorly, even lost the target, because of

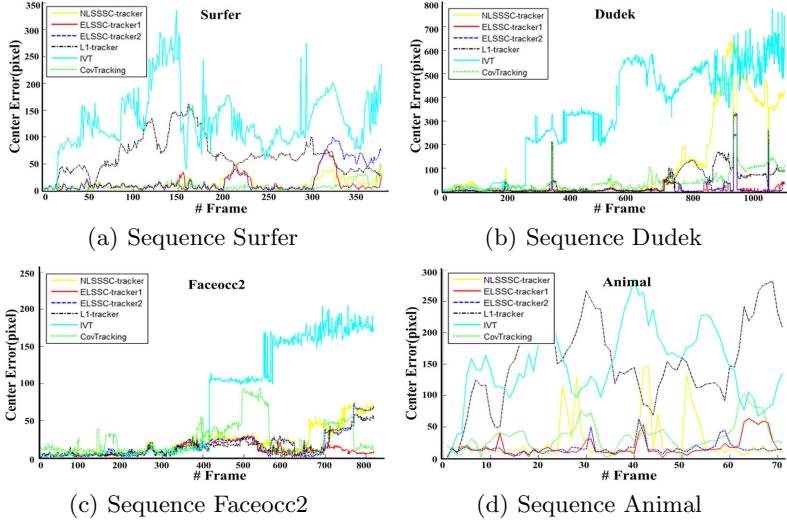


Fig. 4. Quantitative Evaluation in terms of Center Location Error (in pixel)

the number of positive and negative samples are limited(in consideration of the learning efficiency), and incremental updating of classifier in IVT is less effective; CovTracking has large size of candidates(with the definition of integral image, feature extraction of these candidates is so fast, and the costs of which can be ignored), which makes it robust for occlusion, scale variation and blur; NLSSSC-tracker and our two ELSSC-trackers work well when the targets are occluded, especially our two trackers.

For motion blur, our two trackers work better than IVT and original ℓ_1 -tracker, moreover, CovTracking also reveals its ability for blur, see #4, #9 and #38 in Fig.3(d). The animal runs and jumps fast(motion blur) with splashing a lot of water splashing(occlusion). IVT and ℓ_1 -tracker fail both from #4, and never recover. Our original and improved trackers lost target at #31 and #41, and recover at #33 and #44(Fig.3(d)). At #12~#21 and #44~#71, improved ELSSC-tracker works better than other trackers.

For rotation and scale variation, our trackers still work robustly, see Fig.3(a) and 3(c), Fig.4(a) and 4(c). The surfer staggers forward and back in Surfer sequence and the man turns left, turns right and occludes by book in Faceocc2 sequence, five trackers except IVT perform nice for these challenges.

Tab.1 summarizes the average frames per second, improved ELSSC-tracker works much faster than the other trackers; IVT, proposed by Ross[8], is faster than improved ELSSC-tracker when dealing with Surfer and Dudek sequences, but the success rates of IVT is much worse than improved ELSSC-tracker. It is sensitive, when the occlusion, rotation, motion blur of target is appeared in tracking; ℓ_1 -tracker [4] performances good in most frames, but also time-consuming even fail sometimes; Cov-Tracking is suitable for occlusion and rotation.

Table 1. Average Frames per Second. The best two results are shown in red and blue.

Method Video \	IVT	CovTrack	ℓ_1 -tracker	NLSSC	ELSSC1	ELSSC2
Surfer	2.8694	1.5707	0.0358	0.0141	0.0156	2.3469
Dudek	3.3211	1.2454	0.0388	0.0171	0.0179	3.2454
Faceocc2	2.7886	1.1278	0.0180	0.0107	0.0142	3.1278
Animal	1.8979	1.2534	0.0312	0.0150	0.0071	3.2534

5 Conclusion

In this paper, because sparsity and stability cannot be reached simultaneously, in ℓ_1 -problem[5][6] and the high time-consuming of NLSSC-tracker[6], we propose a novel ELSSC-tracker for efficient tracking with Euclidean local-structural(ELS) constraint. Our tracker can be considered as a ℓ_1 -tracker with a reconstructed over-completing dictionary which is different from that in original ℓ_1 -tracker and NLSSC-tracker. We also simplify the high-scale ℓ_1 -problem in original ELSSC-tracker into a much smaller one in our improved ELSSC-tracker. Experiments demonstrate the sparsity, stability and efficiency of our tracker.

But there are also some aspects required to be studied in future, including:(1)the convergence of our two ELSSC-trackers proposed in this paper need to be proved strictly; (2)the general index for quantitative and qualitative analysis of sparsity and stability is needed, not only for ℓ_1 , NLSSC- and our two ELSSC-trackers, but also for other stable ℓ_1 -regularized minimization used in pattern classification, data reconstruction, etc.

References

1. Zhang, S.P., Yao, H.X., Sun, X., et al.: Sparse coding based visual tracking: Review and experimental comparison. *Pattern Recognition* 46(7), 1772–1788 (2013)
2. Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A Survey. *Acm Computing Surveys (CSUR)* 38(4), 1–45 (2006)
3. Candès, E.J., Wakin, M.B.: An Introduction to Compressive Sampling. *IEEE Signal Processing Magazine* 25(2), 21–30 (2008)
4. Mei, X., Ling, H.B.: Robust Visual Tracking and Vehicle Classification via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(11), 2259–2272 (2011)
5. Xu, H., Caramanis, C., Mannor, S.: Sparse Algorithms Are Not Stable: A No-Free-Lunch Theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(1), 187–193 (2012)
6. Lu, X.Q., Yuan, Y.Y., PingkunLu, X., et al.: Robust visual tracking with discriminative sparse learning. *Pattern Recognition* 46(7), 1762–1771 (2013)
7. Daubechies, I., Defrise, M., Mol, C.: An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. *Communications on Pure and Applied Mathematics* 57, 1413–1457 (2004)

8. Ross, D.A., Lim, J., Lin, R.S.: Incremental learning for robust visual tracking. *International Journal of Computer Vision* 77, 125–141 (2008)
9. Porikli, F., Tuzel, O.: Covariance Tracking using Model Update Based on Lie Algebra. In: 2006 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2006 (2006)
10. Zhang, J., Wang, H.Y.: A Fast Sparse Coding Algorithm for Video Object Tracking. *Journal of Frontiers of Computer Science and Technology* 6(8), 760–768 (2012)
11. Everingham, M., Gool, L.V., Williams, C., et al.: The Pascal Visual Object Classes(VOC) Challenge. *IJCV* 88(2), 303–338 (2010)

An Efficient Semi-supervised Hashing Method Based on Graph Transduction

Xiumei Wang, Xianjun Gao, Jie Li, and Ying Wang

School of Electronic Engineering, Xidian University,
Xi'an 710071, China
wangxiumei@gmail.com

Abstract. Hashing-based method has been widely used as the approximate nearest neighbor search model. Kernel based hashing method is a representative hashing-based method because it can make full use of supervised information. It can achieve good performance under the case that the supervised information is sufficient. However, its performance would be influenced with limited label information. In order to deal with this case, we propose a novel semi-supervised hashing method based on label propagation. The proposed method first transfers the labeled information based on graph transduction from the label samples to unlabeled samples, and then constructs a new label matrix. The proposed method can efficiently use the label information to further improve its performance. Extensive experimental results verify the validity of the proposed method.

Keywords: hashing, graph transduction, nearest neighbor search, binary codes.

1 Introduction

In recent years, with the development of the Internet, web data are increasing rapidly. On the one hand, the gigantic datasets offer the opportunity for machine learning and pattern recognition; on the other hand, the adapted methods will be required for dealing with the large datasets. Among the adapted methods, approximate nearest neighbor (ANN)[1] search is an important method because it can find the similarity of data samples. For example, if we want to query an image from the web source, a fast ANN method for measuring the similarity will accelerate the query speed and save the query time.

Hashing-based method is a kind of popular ANN search model which converts each data into a binary code, then it implements approximate nearest neighbor search in the Hamming space. The advantage of hashing method is that the binary code is highly compressed, so it can be loaded into the main memory, and the hamming distance can be computed quickly by using XOR operation. There are some basic principles should be followed in designing hashing model[2]. Firstly, similar data samples should be mapped to similar binary codes. Secondly, the binary code can be easily obtained. Finally, the length of binary code should be as short as possible for saving storage.

One of the most widely known method is locality sensitive hashing(LSH)[3]. LSH ensures that the similar samples can be hashed into same bins with high probability of collision. But it requires long binary code to preserve good precision, which need large storage and cost search time. To obtain more compact hash codes, data-dependent hashing methods have been proposed. Instead of generating random projection in LSH, these methods can train projection such as principal component analysis based hashing (PCAHash)[4] and linear discriminant analysis hashing (LDAH)[5]. These methods generate linear projection for hashing function. In order to deal with nonlinear case, some kernel-based hashing methods have been proposed, such as anchor graph hashing (AGH)[6], binary reconstructive embedding (BRE) [7], kernel-based supervised hashing (KSH)[8]. Spectral hashing (SH)[2] is a well-known method which obtains hashing function based on the graph Laplacian rather than assuming linear projection. Among these hashing methods, some are unsupervised methods, such as LSH, SH, AGH, and some are supervised methods, such as LDAH, KSH. For unsupervised methods, only relevance of the data is used to learn hashing function. For supervised methods, we only use the label of the data, which may rely on the number of the given label. Compared with unsupervised methods, supervised methods may reach more accurate results under the condition of enough label information. When provided labels are limited, the experimental results will also be very limited improvements. However, labeling the data samples often requires expensive price. Therefore, semi-supervised hashing method is required in many real world problems, which can utilize the small label information and achieve convincing results. In this paper, we propose a semi-supervised hashing method based on label propagation. It can make full use of the limited label information and reach superior performance than other existing algorithms.

The rest of this paper is organized as follows. The section 2 introduces the hashing function. The section 3 describes the process of label propagation and the proposed semi-supervised hashing method. The section 4 shows the experimental results of our semi-supervised hashing method and some other representative hashing methods. The section 5 concludes this paper and points out the further researches.

2 Hashing Function

Given a database $X = \{x_i\}_{i=1}^n$ with n samples, each sample x_i is a d -dimensional data, i.e., $x_i \in R^d$. The aim of hashing-based methods is to learn a set of hashing functions which map each sample x_i to a binary code $b_i \in \{1,0\}^r$ ($1 \leq i \leq n$) ,

$$H(x) = \{h_k(x)\}_{k=1}^r. \quad (1)$$

The binary code can be obtained by,

$$b_i = H(x_i). \quad (2)$$

Where $i = 1, \dots, n$. Using $B = [b_1, \dots, b_n]^T \in \{1, 0\}^{n \times r}$ to represent the binary code of given dataset, Eq.(2) can be rewritten with matrix. Then the hashing transform can be represented as,

$$B = H(X) \quad (3)$$

3 Graph Transduction Based Semi-supervised Hashing

The supervised method can obtain good results when there is ample label information. In practice, the label information is very limited. This is not enough for training an accurate hashing function. In order to get more label information, we select a label propagation method based on graph transduction, i.e., LGC[9], and then one semi-supervised hashing method will be established based on the label propagation.

Consider the database $X = \{x_i\}_{i=1}^n$ with l_0 labeled data $X_{l_0} = \{x_i\}_{i=1}^{l_0}$. The corresponding label set $L = \{1, \dots, c\}$. The label information of X can be represented with a matrix Y , $y_i \in R^c$, that is, if the label of x_i is d , $1 \leq d \leq c$ then $y_{id} = 1$ and the other elements of y_i are equal to 0. If x_i is unlabeled sample, all the elements of y_i are equal to 0. To inference more label information, LGC is used for label propagation. The process of label propagation can be concluded as follows,

- (1) Computing similarity of all samples,

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \quad (4)$$

- (2) Constructing the neighborhood graph;
- (3) Labeling the unlabeled samples according to the graph transduction.

With the help of the graph transduction, we can inference the label information of the unlabeled samples. Therefore, enough labeled samples will be obtained for training the hashing function. The hashing function may be linear or nonlinear. In some case, the linear function cannot obtain ideal results, and then nonlinear hashing functions have been given. The kernel-based supervised hashing is a representative method[8]. KSH is a supervised hashing method because it uses the label information in ANN search.

Assumption we get l labeled samples by using label propagation, the label matrix $S \in R^{l \times l}$ can be defined as,

$$S_{ij} = \begin{cases} 1, & (x_i, x_j) \in M \\ -1, & (x_i, x_j) \in C \end{cases} \quad (5)$$

Where M denotes the set that sample x_i and x_j are similar or they belong to the same class, that is, y_i is equal to y_j . C denotes the set that sample x_i and x_j are dissimilar or they belong to different classes, that is, y_i does not equal to y_j .

In order to reduce the computation complexity, the next step is getting the m anchor points, $U = \{u_1, \dots, u_m\}$. The anchor can be obtained by k-means clustering[6] or random select method[9]. In the proposed semi-supervised model, k-means is selected to determine anchors. Then the hashing function is represented as,

$$h_i(x) = \text{sgn}(K(x)w_i). \quad (6)$$

Where $w_i = [w_{i1}, \dots, w_{im}]^T$ is the projection vector which can be obtained from training process. $W = [w_1, \dots, w_r] \in R^{m \times r}$ is the projection matrix. $k(x, u_i)$ with $1 \leq i \leq m$ is the relationship between the data x and the anchor point u_i , then,

$$K(x) = [k(x, u_1), \dots, k(x, u_m)]. \quad (7)$$

$K(x)$ can be regard as a mapping function. That is $K : R^d \mapsto R^m$. Then the binary code of the given database B can be computed,

$$B = \text{sgn}(K(X)W). \quad (8)$$

Where $K(X) \in R^{n \times m}$, $W \in R^{m \times r}$.

One important principal of hashing-based methods is preserving the similarity of the samples. The similarity of the samples can be reflected by the code inner product, which can be defined as,

$$D_c(b_i, b_j) = b_i b_j^T \quad (9)$$

The relationship of code inner product and the Hamming distance can be represented as,

$$D_c(b_i, b_j) = r - 2D_h(b_i, b_j). \quad (10)$$

Where $D_h(b_i, b_j) = \frac{1}{4} \|b_i - b_j\|$, $D_h(b_i, b_j) \in \{0, 1, \dots, r\}$, $D_c(b_i, b_j) \in \{-r, \dots, 0, \dots, r\}$. If the codes of two samples are similar, the value of their Hamming distance will be small and the value of their code inner product will be large. When the labels of x_i and x_j are same with each other, $D_c(b_i, b_j) = r$, $S_{ij} = 1$. The objective function with least-squares will be,

$$\begin{aligned} \min_{W \in R^{m \times r}} Q(W) &= \left\| \frac{1}{r} H(X_l) (H(X_l))^T - S \right\|_F^2 \\ &= \left\| \frac{1}{r} \text{sgn}(K(X_l)W) (\text{sgn}(K(X_l)W))^T - S \right\|_F^2. \end{aligned} \quad (11)$$

The greedy optimization algorithm is used here to solve the optimal value of this objective function, and the sign function is replaced with the sigmoid-shaped function[8].

The whole flowchart of the proposed semi-supervised hashing method is given in Fig.1.

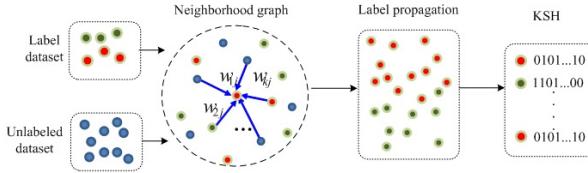


Fig. 1. The flowchart of the semi-supervised hashing method

4 Experiments

In this section, we will evaluate the proposed semi-supervised hashing method on two datasets: the MNIST database of handwritten digits and a subset (9K) of the Photo Tourism (Notre Dame part) image patch dataset. Some popular hashing methods, such as, LSH SH, PCAITQ, KSH, will be presented for comparison. Hamming ranking and Hash lookup will be adopted as evaluation protocols in the following experiments.

4.1 Database

The MNIST[6,11] dataset consists of 70,000 digital images. Each image is a 28· 28 handwritten image. In experiments, each image will be transform to a 784 dimension vector. 69,000 samples are used for training hashing function, and others for testing as in the paper[6].

The Photo Tourism image[11] is taken from Photo Tourism reconstructions from Trevi Fountain (Rome), Notre Dame (Paris) and Half Dome (Yosemite). The dataset consists of 1024· 1024 bitmap images, each containing a 16· 16 array of image patches. In this paper, we use a subset of Notre Dame (Paris) part. This part contains 104,106 image patches with 35,280 classes. We just use a subset of this part with 662 classes, which is the same with the dataset used in Ref.[12].

4.2 Parameter Selection

There are some parameters we have to determine before doing the experiment. In label propagation, an important parameter is α . We fix $\alpha = 0.9$. In Supervised

Hashing, the anchor point number $m = 300$, and k-mean method is used to obtain the anchor point. The values of two parameters are similar to them in LGC and KSH respectively. The label number of MNIST dataset is 100 and 200, and the label number of Photo Tourism dataset is 662. The labeled samples are used to train KSH, and estimate the label of unlabeled samples for the proposed SSH.

4.3 Experiments Results

Fig. 2 show the results of MNIST dataset with 100 labeled data and Fig.3 show the results of MNIST dataset with 200 labeled data. The number of training dataset is very small compared with the number of the whole MNIST dataset.

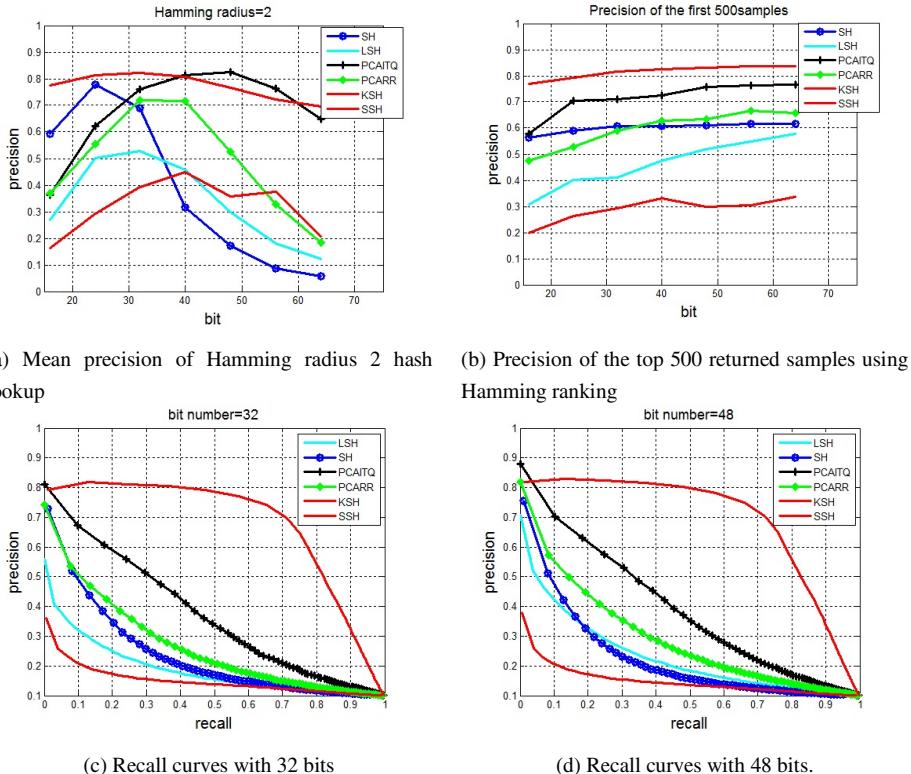


Fig. 2. The results of 100 labeled data on MNIST dataset

Figure 2(a) and Figure 3(a) show the results of precision using hash lookup less than or equal to hamming radius 2. Fig. 2(b) and Fig. 3(b) show the results of precision of the top 500 returned samples using Hamming ranking. The length of binary code is varied from 16 to 64. In two Figures, SSH and PCAITQ methods outperform other methods, and the performance of KSH is limited. When the binary code length

is less than 40, KSH becomes the worst one among these hashing methods. The reason is that the given labeled data is too less to train a powerful hashing function. Fig.2(c), Fig.3(c) and Fig.2(d), Fig.3(d) show the recall curves with 32 and 48 bit binary code. It can also been seen the performance of SSH is well enough from the Figures. We can see from Figure 2 and Figure 3 that our method can perform well even when the code is short.

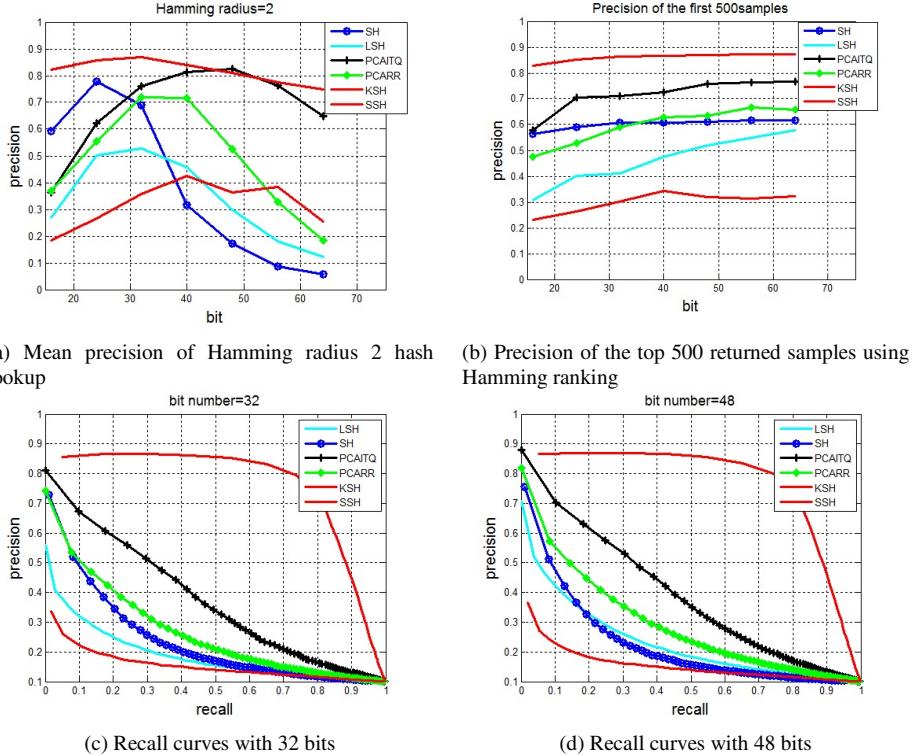


Fig. 3. The results of 200 labeled data on MNIST dataset

Fig. 4 shows the results of Photo Tourism dataset with 662 labeled data. Fig. 4(a) show the results of precision using hash lookup less than or equal to hamming radius 2. Fig. 4(b) show the results of Precision of the top 500 returned samples using Hamming ranking. Fig. 4(c) and Fig. 4(d) show the recall curves with 32 and 48 bit binary code. The result of hash lookup within hamming radius 2 is a little bad in Fig. 4(a). But the hamming ranking curve and recall-precision curve still outperform other hashing methods in other three subfigures. The precision of KSH is better than Fig. 2 and Figure 3 for hamming ranking and recall curves because the labeled data number is larger. But it still cannot achieve better results than SSH.

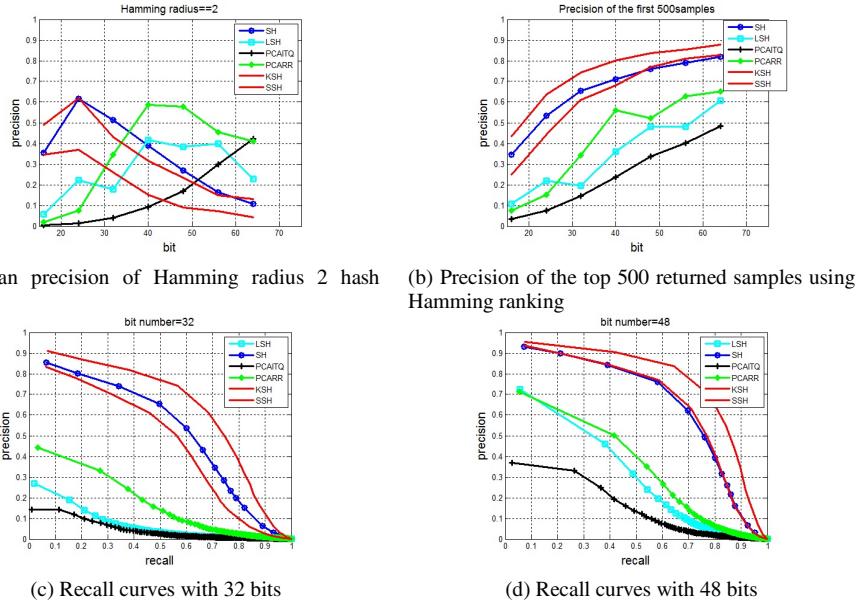


Fig. 4. The results of 662 labeled data on Photo Tourism dataset

5 Conclusion

In this paper, we proposed a novel semi-supervised hashing method based on graph transduction. The proposed method can take full advantage of the label information. We attempt to solve the problem that supervised hashing method cannot obtain good results when labeled samples are very small. Our method empirically outperforms other hashing techniques, especially in cases when the labeled information is very small. In the experimental, the percent of labeled samples in whole dataset is lower than 0.2%, whereas, the results are inspiring.

Acknowledgments. This research was supported partially by the National Natural Science Foundation of China (Grant Nos. 61125204, 61172146, 61100158 and 61201294), by the Fundamental Research Funds for the Central Universities (K50511020002), by the China Postdoctoral Science Foundation under Grant 2012M521746, the Industrial Public Relation Project of Shaanxi Technology Committee (2012K06-40).

References

- Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proc. of ACM Symposium on Theory of Computing, pp. 604–613 (1998)
- Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Proc. of Advances in Neural Information Processing Systems, pp. 1753–1760 (2008)

3. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proc. of 25th International Conference on Very Large Data Bases, pp. 518–529 (1999)
4. Gong, Y., Lazebnik, S.: Iterative quantization: A Procrustean approach to learning binary codes. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 817–824 (2011)
5. Strecha, C., Bronstein, A., Bronstein, M., Fua, P.: LDAHash: Improved Matching with Smaller Descriptors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 34(1), 66–78 (2012)
6. Liu, W., Wang, J., Kumar, S., Chang, S.: Hashing with Graphs. In: Proc. of International Conference on Machine Learning, pp. 1–8 (2011)
7. Kulis, B., Darrell, T.: Learning to Hash with Binary Reconstructive Embeddings. In: Proc. of Advances in Neural Information Processing Systems, pp. 1042–1050 (2009)
8. Liu, W., Wang, J., Ji, R., Jiang, Y., Chang, S.: Supervised Hashing with Kernels. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2074–2081 (2012)
9. Zhou, D., Bousquet, O., Lal, T., Weston, J., Scholkopf, B.: Learning with local and global consistency. In: Proc. of Advances in Neural Information Processing Systems, pp. 321–328 (2004)
10. <http://yann.lecun.com/exdb/mnist/>
11. <http://phototour.cs.washington.edu/patches/default.htm>
12. <http://www.ee.columbia.edu/ln/dvmm/downloads/WeiKSHCode/dlform.htm>

Fuzzy-PI Switch Control in Intermediate Frequency Heating Process of 3PE-Coating^{*}

Xu Yang^{1,2,**}, Bo Wen^{1,2}, Chao-nan Tong^{1,2}, Yu-zhi Feng³, and Yan Zeng^{1,2}

¹ School of Automation & Electrical Engineering,

Key Laboratory of Advanced Control of Iron and Steel Process (Ministry of Education),
University of Science and Technology Beijing, Beijing 100083, China

² School of Automation & Electrical Engineering,

University of Science and Technology Beijing, Beijing 100083, China
³ Beijing Institute of Astronautical Systems Engineering, Beijing, 100076, China
yangxu@ustb.edu.cn, {gzwebber, bjyzfeng}@qq.com,
tcn@ies.ustb.edu.cn, zengyan0612@126.com

Abstract. Three layer PE coating (3PE), which combines the advantages of epoxy powder and polyethylene coating, is an excellent approach of anti - corrosion in the long pipeline of natural gas and refined oil. Steel pipe heating temperature is one of the key parameter that determines the performance of the anti-corrosion coating. The model of heating process is established in this paper. However, the large inertia and strong disturbance of the system make it difficult to get satisfactory control result with conventional PI controller. A switch fuzzy controller was proposed in view of this situation. By comparing with the PI controller, the new controller can not only reduce the response time, but also eliminate the overshoot. A simulation with actual production data in the heating process of a 3PE anti-corrosion factory production line shows the validity of the model and the switch controller.

Keywords: 3PE-coating, Intermediate Frequency Heating, Fuzzy control, Switch control.

1 Introduction

As one of the most advanced pipeline anti-corrosion technology, 3PE anti - corrosion structure is widely used in steel pipe production of oil and gas pipelines [1,2]. Production process includes shot blasting, coating, groove, and coating process is further divided into heating, dusting, winding and cooling.

After pretreatment, steel pipes are conveyed to the three-layer coating process, in which 60-100um FBE, 250-400 um adhesive and 1.5-3.0 mm PE are coated layer by layer. 3PE coating layer has outstanding resistance to mechanical damage and has excellent adhesive properties [3,4].

* **Foundation Item:** Item Sponsored by National Natural Science Foundation of China (51205018), China Postdoctoral Science Foundation Funded Project (2012M510321), and Fundamental Research Funds for the Central Universities (FRF-TP-12-104A, FRF-SD-12-008B).

** Corresponding author.

Some scholars had studied the system, but the large inertia and strong disturbance still make it difficult to establish an accurate model of the heating process [5]. In this paper, the characteristics of this process are studied, and the heating and cooling model is improved based on actual data. We further build a control system structure, and propose a fuzzy-PI switch controller, with which the response time of the system is reduced and the overshoot is eliminated as well.

2 Heating and Heat Dissipation Model

2.1 Process and Model

The heating temperature of the steel pipe determines the FBE gel time, which is an important factor affects the quality of production. The heating process model should firstly established, and then a temperature controller be designed, which can make the temperature of the steel pipe follow the given value quickly and smoothly, in order to improve the quality of production.

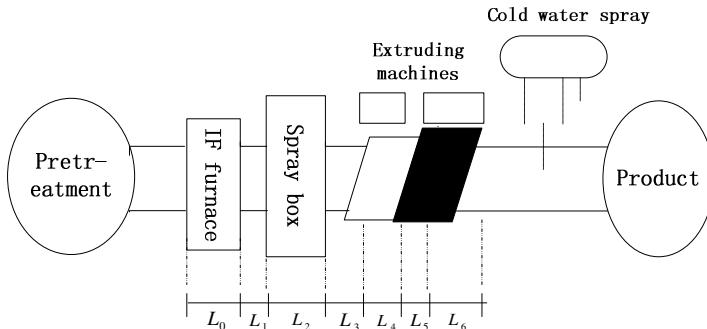


Fig. 1. The coating segment process

In heating process, after rubber wheels drive pipe axial forward into the intermediate frequency (IF) furnace by friction, the IF furnace then heat the tube to the expected temperature by the principle of electromagnetic induction. Consider the previous temperature of the pipe r_0 and environment temperature r_1 , the model is shown in figure 2.

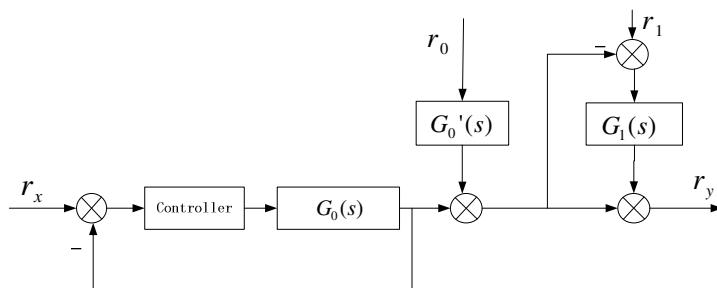


Fig. 2. Heating and heat dissipation model

As we can see from [5], the pipe heating process can be modeled as a first-order inertia, which is shown below:

$$G_0(s) = \frac{K}{(T_0 s + 1)} \quad (1)$$

where T_0 is the time constant of heating.

$$K = \frac{r_y}{U_{dc}} \quad (2)$$

K is defined as the static gain from average from U_{dc} to r_y , U_{dc} denotes the average output dc voltage of the rectifier bridge and r_y is the temperature of steel pipe at IF furnace outlet, the unit is K/V .

$$G'_0(s) = \frac{1}{(T_0 s + 1)} \quad (3)$$

The relationship of T_0 , L_0 , v_0 can be expressed as below:

$$T_0 = \frac{L_0}{v_0} \quad (4)$$

$$G_1(s) = \frac{K}{(T_1 s + 1)} \quad (5)$$

T_1 is the time constant of heat dissipation.

2.2 Model Improvement

The steel pretreatment will increase the temperature of the steel pipe before heating up, in order to eliminate the steel pipe temperature on the frequency heating system, detect and join the feed forward block weighs $-1/K$. After frequency heating, the steel pipe would go through some distance before reaching the extruder. During this period, pipe temperature will decrease by the environmental impact. In order to represent the original system more accurately, we improve system model as shown in figure3.

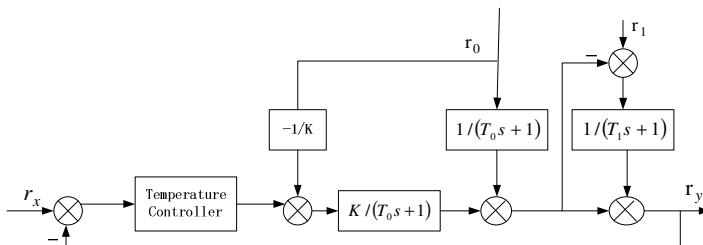


Fig. 3. System structure of heating process

3 Switch Controller Design

3.1 Fuzzy Logic Controller (FLC)

In actual production, due to the power limit of the intermediate frequency furnace, the heating process is relatively slow.

In order to meet the production requirements, we typically use the maximum power for heating. There will be a large overshoot using the conventional PI controller. However, steel pipes cooling process takes a long time. The controller should be designed to avoid the emergence of large overshoot, which means PI is inappropriate in controlling this model. It has been proved that the fuzzy controller perform for nonlinear, complex and time-vary system, especially for some systems without accurate mathematical models [6-8]. Under this situation, we consider to apply fuzzy approach to the system design, which can increase the damping of the system and decrease overshoot.

Fuzzy controller's input variables and output variables are given in the form of fuzzy set. We often take the deviation e and the derivative of deviation de as input variables and the amount of control u as output variable. In this paper, $e = r_y - r_x$, u is supply voltage of IF heating furnace.

The seven fuzzy subsets collection are commonly used.

{NB, NM, NS, ZO, PS, PM, PB}

NB=(Negative Big)

NM=(Negative Medium)

NS=(Negative Small)

ZO=(Zero)

PS=(Positive Small)

PM=(Positive Medium)

PB=(Positive Big)

The quantization level of the language variables were taken 13, i.e. $E = de = U = \{-6 - 5 - 4 - 3 - 2 - 1 0, 1, 2, 3, 4, 5, 6\}$.[18] The basic domain of the error e [-20 °C, 20 °C]; basic domain of error change def[- 20 °C / dec, 20 °C / dec]; basic domain of control output u [-30 V, 30 V] quantitative factor $k_e = 0.3$, $k_c = 0.3$, $k_u = 5$.

Imitate the human-control behavior, the membership function should be normal fuzzy variables. Sharper the membership function curve seems, a higher resolution it will have; on the contrary, if the membership function shape is relatively flat, the control features will be more gentle and stable [9]. Therefore, in choosing the membership function of the fuzzy variables we should follow this principle: use fuzzy set of low-resolution in larger error area, and high-resolution fuzzy sets in smaller

error area, in order to achieve high control precision and good stability control effect at the same time. The membership functions are shown in figure 4 below.

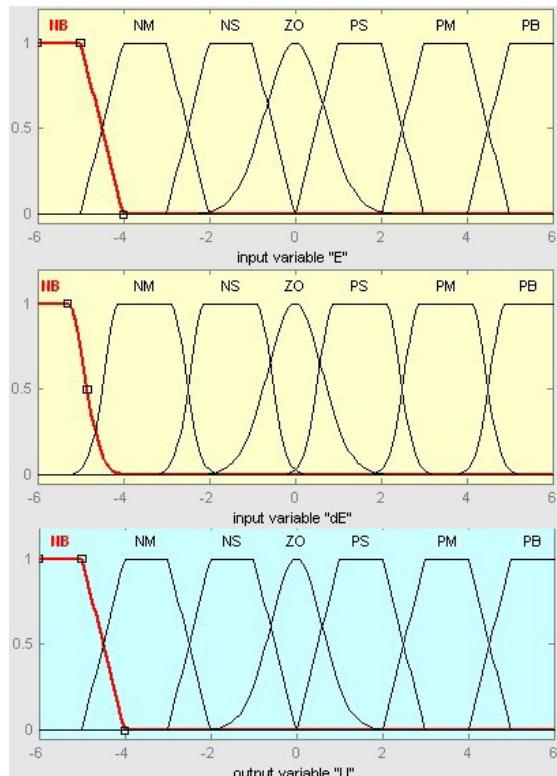


Fig. 4. Membership function curve of “E” ”dE” and “U”

Control rules is shown in the following table which represents 49 ‘if ... then ...’ sentences. For example: If E is NB and DE is PS then U is PM. Use Mamdani’s decision-making method in Fuzzy decision and gravity center method and in defuzzification.

Table 1. Fuzzy control rules table

dE / E	NB	NM	NS	ZO	PS	PM	PB
NB	PB	PB	PB	PB	PM	ZO	ZO
NM	PB	PB	PB	PB	PM	ZO	ZO
NS	PM	PM	PM	PM	ZO	NS	NS
ZO	PM	PM	PS	ZO	NS	NM	NM
PS	PS	PS	ZO	NM	NM	NM	NM
PM	ZO	ZO	NM	NB	NB	NB	NB
PB	ZO	ZO	NM	NB	NB	NB	NB

An switch controller can be designed based on the FLC above. Since the fuzzy controller lacks integral part, its ultimate control effect is similar to the PD controller. There is steady-state error, a slight shock may also occur near the equilibrium point. Parallel an integrator besides the fuzzy controller, the hybrid controller won't generate steady-state error [10,11].

However, hybrid fuzzy controller doesn't react as fast as PI controller in large error area. To solve this problem, we propose a strategy shown in figure 4. When the deviation is greater than a certain threshold, PI controller is applied, in order to speed up the system responding. Once the deviation is smaller than the threshold, the switch fuzzy-PI controller is applied to increase damping of the system and reduce or eliminate the overshoot.

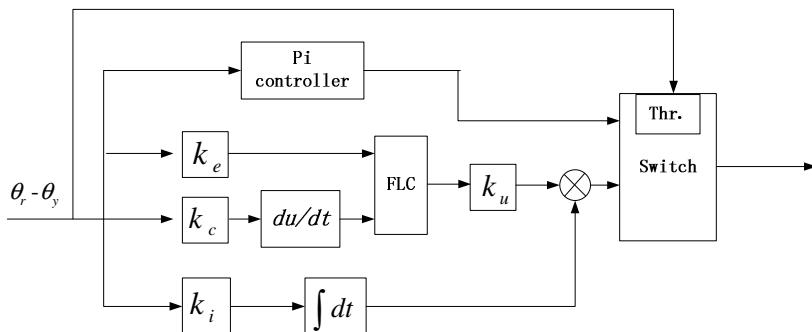


Fig. 5. Switch controller structure

4 Simulation Verification and Result Analysis

In order to verify the validity of the model and controller above, a simulation verification analysis was carried out by using ø 1016 pipe parameters and actual production data in an anti-corrosion steel pipe factory, using Matlab/Simulink platform to do simulation verification analysis. The axial velocity of the pipe is 1.67 m/min, the expected heating temperature r is 200 ° C. Due to the impact of pretreatment part and the thermal conductivity, assume that the steel pipe temperature is 40 ° C before entering the IF furnace. The environmental temperature is 20°C. According to the on-site production data, take $T_0 = 35.9$ s, $T_f = 800$ s. The output control table of fuzzy controller is built and calculated by the Fuzzy toolbox.

Step signal is given at t=1s. Blue line in Figure.6 is the PI control curve, a fast initial corresponding can be seen, but the overshoot is very large which should be avoided in heating process. The blue line is the fuzzy PI controller curve, the result shows that the system damping is significantly increased and the overall response time is shortened compared to that of the PI controller. The red line is the switch fuzzy controller curve, the new controller combines the advantages of the first two control methods and can complete the regulation within 15s with elimination of the overshoot, which indicates a better performance over the PI controller and fuzzy-PI controller.

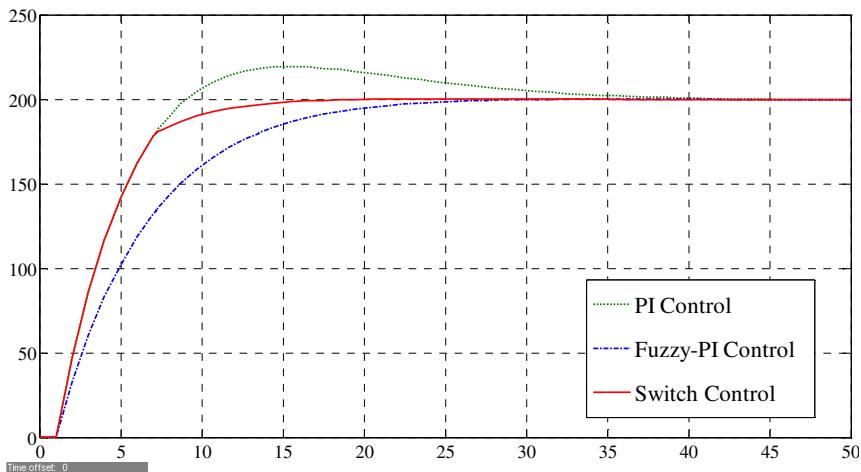


Fig. 6. Response curve of fuzzy-PI switch controller

5 Conclusion

The heating temperature of steel pipe is a key parameter of 3PE production that determines the performance of the anti-corrosion coat. According to the heating process of 3PE-coating production, the heating and heat dissipation model is improved. Furthermore, this paper proposed a switch controller based on fuzzy theory. The simulation experiment is built based on the heating process model and the actual production data of a 3PE anti-corrosion factory production line. A comparison with PI and fuzzy-PI controller shows that the new controller has faster convergence speed without overshoot, which verifies the validity of the switch control strategy.

References

1. Gong, S.M.: Application of three layer PE in China. *J. Anticorrosion & Insulation technology* 12(4), 11–13 (2004)
2. Ying, G.Y., Lei, S.L.: Practice of 3PE pipeline anticorrosion in China. *J. Welded Pipe* 30(1), 8–11 (2007)
3. Zheng, M.G.: The three PE covering layer system of pipeline anticorrosion. *J. Foreign Oilfield Engineering* 31(6), 41–46 (2000)
4. Feng, H.Q.: Development and application of pipeline anticorrosion coating technology of the three layer PE. *J. Welded Pipe* 22(3), 18–21 (1999)
5. Li, W., Wang, J., Qiao, P.Y.: Study on Theory Modeling Method About Steel Tube Transferring Heat in 3-Layer Pecoated Pipe Process. *J. Journal of Gansu Sciences* 16(1), 76–80 (2004)

6. Baghramian, A., Ghorbani, E.: Fuzzy Controller of Luo Converter for Controlling of DC Motors Speed. C. Power Electronics. In: Drive Systems and Technologies Conference (PEDSTC), pp. 170–175 (2013)
7. Hanafi, D.: The Quarter Car Fuzzy Controller Design Based on Model from Intelligent System Identification. J. Industrial Electronics & Applications 2009(2), 930–933 (2009)
8. Patel, A.V., Mohan, B.M.: Analytical structures and analysis of the simplest fuzzy PI controllers. Automatica 38, 981–993 (2002)
9. Wang, J.F., Lu, Z.D.: Determining method of the membership function of fuzzy control. J. Henan Science 18(4), 22–25 (2000)
10. Tao, Y.H., Yin, Y.X., Ge, L.S.: The new PID control and its application. Mechanical Industry Press, M. Beijing (1998)
11. Lee, J.: On methods for improving performance of PI-type fuzzy logic controllers. J. IEEE Transactions on Fuzzy Systems, 298–301 (1993)

An Improved Method in Change Detection of Multitemporal Remote Sensing Image

Fangshun Liao^{1,2}, Sufen Yu², Ying Li^{1,2}, and Yanning Zhang^{1,2}

¹ Shaanxi Provincial Key Lab. of Speech & Image Information Processing,
Northwestern Polytechnical University. Xi'an, 710072, Shaanxi, China

² Science and Technology on Electro-optic Control Laboratory, Louyang 471009, China
lfsgs30@gmail.com, sfystcl@163.com,
{lybyp, ynzhang}@nwpu.edu.cn

Abstract. Traditional Markov random Field (MRF) methods assume that neighboring pixels tend to have the same label. However, this assumption is always inconsistent with the actual situation and affects the resultant accuracy of the algorithm. To overcome this, we propose an object-based Markov Random Field (OMRF) model and a change detection method based on OMRF model. The OMRF model assumes that pixels within same object are in the same class. First, we generate the difference image from multi-temporal remote sensing images. Second, Mean Shift is applied to extract objects from difference image. Finally, change detection map is generated by iterative algorithm. The experimental results show that the algorithm can effectively improve the detection accuracy of the algorithm on real remote sensing datasets.

Keywords: Change Detection, Remote Sensing Images, Markov Random Field, Object-based.

1 Introduction

Change detection is a process that identifies the difference between two multiple images acquired over the same geographical area at different times. Turgay Celik[1] combined with genetic algorithm, detected change regions and unchanged regions in an optimizing way. In [2], double threshold and semi-supervised support vector machine (S^3VM) algorithm are used to segment the difference image and get the change detection map respectively.

However, the results from the methods above are easily affected by noise. In order to restrain the influence of noise, [3] used multi-resolution and image segmentation method for change detection. In [4-5], methods based on Markov random field (MRF) model are devised for change detection by considering the neighborhood spatial-contextual information of each pixel. But in these MRF model methods, the global consistency assumption that pixel and its neighborhoods are same category in the MRF methods is often not in line with the actual situation and affects the accuracy of the results. Furthermore, MRF model based on neighborhood system generally need to consider the anisotropy of region boundary and the isotropic within the region,

resulting in introducing the spatial-contextual difficultly[6]. [7] considered the anisotropy of the edge with the parameters differences at the boundary and within regions of the MRF model. But, the method does not take full advantage of the information between the pixels within the area.

According to the shortcoming of MRF model above, we present a novel change detection algorithm of remote sensing image based on object-based MRF (OMRF) model. The OMRF model can consider the boundary anisotropy and use the spatial-contextual information greatly. Experimental results demonstrate the proposed approach is superior to the MRF method.

2 Image Model and Problem Formulation

Consider two multi-temporal remote sensing images, X_1 and X_2 of size $M \times N$ acquired at the same geographical area at two different times, t_1 and t_2 . We assume that images have been registered and corrected. D represents the difference image.

$$D = \{D(i, j), 1 \leq i \leq M, 1 \leq j \leq N\} \quad (1)$$

where $D(i, j)$ is computed as follow

$$D(i, j) = |X_1(i, j) - X_2(i, j)|, 1 \leq i \leq M, 1 \leq j \leq N \quad (2)$$

Let $\Omega = \{\omega_n, \omega_c\}$ denote the unchanged class and change class. And $C = \{C_l, 1 \leq l \leq 2^{M \times N}\}$ depicts all label set, where $C_l = \{C_l(i, j), 1 \leq i \leq M, 1 \leq j \leq N\}$, with $C_l(i, j) \in \Omega$. According to the statistical theory and Bayesian formulation, the MRF considers change detection as an optimal problem which maximizes the posterior probability $P(C_l / D)$ [4]

$$C_k = \arg \max_{C_l} \{P(C_l / D)\} = \arg \max_{C_l} \{P(C_l) p(D / C_l)\} \quad (3)$$

where $P(C_l)$ is the prior model for the class labels, and $P(D / C_l)$ is the joint density function of pixel values in the difference image given by the set of labels C_l [4]. As a further simplification of the problem, we assume the following conditional independence

$$p(D / C_l) = \prod_{\forall D(i, j)} p(D(i, j) / C_l(i, j)) \quad (4)$$

We assume C_l is a Markov random field, so, $P(C_l(i, j))$ can be compute as:

$$\begin{aligned}
& P(C_l(i, j) / \{C_l(g, h), (g, h) \neq (i, j)\}) \\
&= P(C_l(i, j) / \{C_l(g, h), (g, h) \in N(i, j)\}) \\
&= \frac{1}{Z} \exp(-U(C_l(i, j) / \{C_l(g, h), (g, h) \in N(i, j)\}))
\end{aligned} \tag{5}$$

where $N(i, j) = \{(i, j) + (\nu, h), (\nu, h) \in N\}$ with $N = \{(\pm 1, 0), (0, \pm 1), (1, \pm 1), (-1, \pm 1)\}$, is neighbor system of the pixel (i, j) . $U(\cdot)$ is the Gibbs energy function, and Z is a normalizing factor.

$$U(C_l(i, j) / \{C_l(g, h), (g, h) \in N(i, j)\}) = \sum_{(g, h) \in N(i, j)} \beta \cdot \delta(C_l(i, j), C_l(g, h)) \tag{6}$$

$\delta(\cdot)$ is the Kronecker delta function, which be expressed as

$$\delta(C_l(i, j), C_l(g, h)) = \begin{cases} -1, & \text{if } C_l(i, j) = C_l(g, h) \\ 0, & \text{if } C_l(i, j) \neq C_l(g, h) \end{cases} \tag{7}$$

and β is a constant.

In terms of the Markovian approach, maximized (3) is equivalent to the minimization of the follow energy function

$$U(D, C_l) = \sum_{\forall(i, j)} U_{data}(D(i, j) / C_l(i, j)) + U_{context}(C_l(i, j) / \{C_l(g, h), (g, h) \in N(i, j)\}) \tag{8}$$

the energy term $U_{context}(\cdot)$ describes the inter-pixel class dependence, which is determined by (6). $U_{data}(\cdot)$ represents the statistics of the gray levels in the difference image. In the Gaussian model, $U_{data}(\cdot)$ can be computed as

$$U_{data}(D(i, j) / C(i, j)) = \frac{1}{2} \ln 2\pi\sigma_{C_l(i, j)}^2 + \frac{1}{2} (D(i, j) - \bar{\alpha}_{C_l(i, j)})^2 [\sigma_{C_l(i, j)}^2]^{-1} \tag{9}$$

where $\sigma_{C_l(i, j)}^2 \in \{\sigma_n^2, \sigma_c^2\}$ and $\bar{\alpha}_{C_l} \in \{\bar{\alpha}_n, \bar{\alpha}_c\}$ are estimates obtained by the EM[8]algorithm.

3 Change Detection Based on OMRF Model

In the traditional based-MRF model methods, they assume that a pixel belongs to the class if this pixel neighborhoods belong to the class[5]. However, the traditional change detection methods based on MRF do not take into account the anisotropy of the edge, and cannot take advantage of all the relationship between the pixels based on second-order neighborhood system. According to this, we propose OMRF model which assume pixels within the same object belong to same class, and change

detection algorithm based on the OMRF model. So, the proposed algorithm must get the objects. In our paper, Mean Shift algorithm is used.

3.1 Mean Shift Segmentation Algorithm

We extract objects from the difference image by using mean shift segmentation algorithm[9]. The mean shift algorithm is a nonparametric clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. Let x_i , $i=1,\dots,n$ be the d-dimensional input image pixels in the joint spatial-range domain, and the mean shift is

$$m_h(x) = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \quad (10)$$

where

$$g(y) = \begin{cases} e^{-\|y\|^2}, & \text{if } \|y\| \leq 1 \\ 0, & \text{others} \end{cases} \quad (11)$$

And h can be regarded as: the spatial hs defining the spatial window and the spectral hr defining the spectral window. Let MC be a maximum label value of segmentation image L , and we can get object set $SC = \{SC_k, 1 \leq k \leq MC\}$, with $SC_k = \{(i, j) | L(i, j) = k\}$.

3.2 OMRF Model and Algorithm Implements

Traditional based-MRF model method is based on second-order neighborhood system, $P(C_l(i, j))$ determined by the pixels of the (i, j) second-order neighborhood system. In the OMRF model, $P(C_l(i, j))$ determined by the pixel belonging to the same object of the pixel (i, j) , so the conditional probability is computed as follows

$$\begin{aligned} & P(C_l(i, j) / \{C_l(g, h), (g, h) \neq (i, j)\}) \\ &= P(C_l(i, j) / \{C_l(g, h), (g, h) \in SC_k \cap (i, j) \in SC_k\}) \\ &= \frac{1}{Z} \exp(-U(C_l(i, j) / \{C_l(g, h), (g, h) \in SC_k \cap (i, j) \in SC_k\})) \end{aligned} \quad (12)$$

where $U(\cdot)$ is Gibbs energy equation, Z is a normalization parameter.

$$U(C_l(i,j)/\{C_l(g,h), (g,h) \in SC_k \cap (i,j) \in SC_k\}) = \sum_{(g,h) \in SC_k} \beta \cdot \lambda \cdot \delta(C_l(i,j), C_l(g,h)) \quad (13)$$

where λ is a parameter and is computed by $1/\|SC_k\|$. $\|SC_k\|$ is the count number pixels of the object SC_k . We get the change detection map by minimizing the energy function

$$U(D_l C_l) = \sum_{\forall(i,j)} U_{data}(D_l(i,j)/C_l(i,j)) + U_{context}(C_l(i,j)/\{C_l(g,h), (g,h) \in SC_k \cap (i,j) \in SC_k\}) \quad (14)$$

$U_{context}(\cdot)$ determined by (13), and assuming that each random variable in the observed field to meet independent Gaussian distribution, so, we can get the $U_{data}(\cdot)$ by (9).

Under normal circumstances, minimize the energy equation using iterative algorithm. ICM algorithm[10] is used in this paper, which is simple and fast, and has proven to be able to converge to a local minimum (see [10] for more details).

3.3 Overall Algorithm

The proposed algorithm process can be summarized as 3 steps:

1. We get difference image from two remote sensing images which are registered and corrected at the same geographical area at two different times.
2. Using the EM algorithm to estimate the parameters of the Gaussian model and Mean Shift algorithm for image segmentation to get each object.
3. Minimizing OMRF model equations (14) to get change detection map.

The algorithm flow chart is shown in Fig.1.

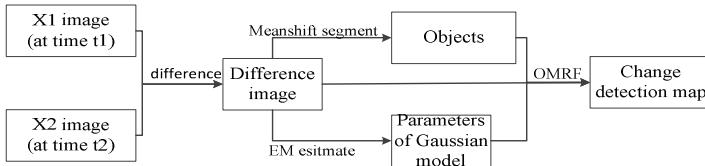


Fig. 1. OMRF algorithm flow chart

4 Experimental Results

In order to verify the effectiveness of the proposed algorithm, we did two different experiments on two data sets. And we implement the change detection method of Lorenzo Bruzzone[4] and the method in [3]. The same set of parameters used is as presented in [4] and [3].

The first data set was acquired over Alaska by Landsat 5 Thematic Mapper (TM) in July 1985 and July 2005. The images with 475·449 pixels as shown in Fig.2 (a) and (b) are chosen for our experiments. The ground truth of the change detection map is shown in Fig.2 (c). The parameters $hs = 7$, $hr = 6.5$ are selected for Mean Shift algorithm, and the parameter $\beta = 1.1$ is set in OMRF. The experimental results are shown in Fig.3.

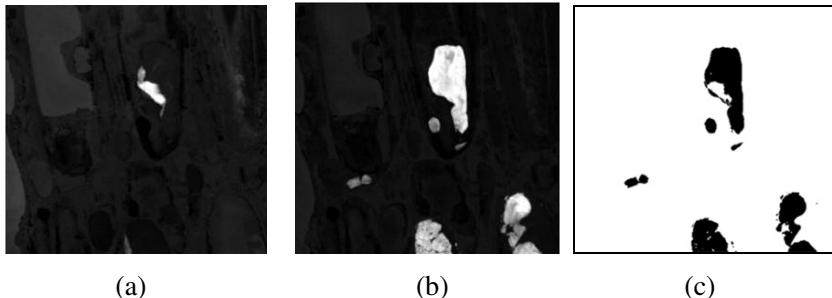


Fig. 2. “Alaska”: image acquired in(a) July 1985 and (b) July 2005 (c)ground truth

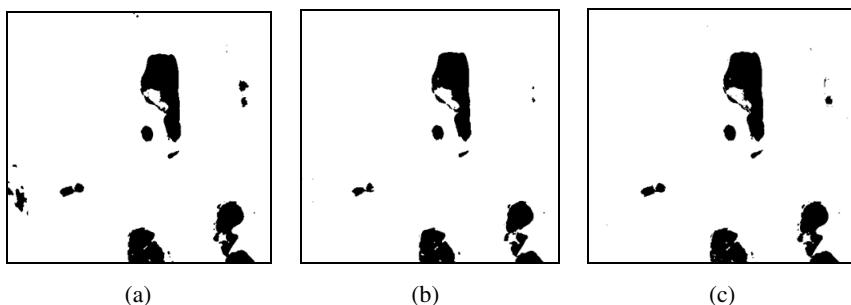


Fig. 3. “Alaska”: change detection maps (a) MRF[4] (b) UDWT-AC[3] (c) OMRF

From the experimental results shown in Fig.3, the proposed method can constrain noise and maintain the boundary of the image. As can be seen from Fig.3.(a) left region, a lot of false pixels was detected by using MRF [4] method. But the pixels were not detected in our approach. The main reason is that MRF model cannot fully take advantage of the spatial relationship of pixels, resulting in the same region pixels into two categories. The results of objective evaluation are tabulated in Table 1.

Table 1. “Alaska”: comparison between different methods in objective evaluation

Method	Miss Detections		False Alarms		Total Errors	
	Pixels	%	Pixels	%	Pixels	%
MRF[4]	34	0.2322	1243	0.6258	1277	0.5988
UDWT-AC[3]	675	4.6103	193	0.0972	868	0.4070
OMRF	12	0.0820	186	0.0936	198	0.0928

The second data set was acquired over Lake Mead in June 1991 , by Landsat7, and May 2000, by Landsat5. The images with 316· 351 pixels as shown in Fig.4 (a) and (b) are chosen for our experiments. The ground truth of the change detection map is shown in Fig.4 (c). The parameters of $hs = 7$ and $hr = 10.5$ are chosen for our simulation experiments. And the parameter of $\beta = 1.1$ is selected for OMRF. The change detection results by using the implemented and proposed methods shown Fig.5, and are tabulated in Table 2. The miss detection and false alarm of our method are both less than UDWT-AC[3] method, which shows that we can define the boundary more accurate.

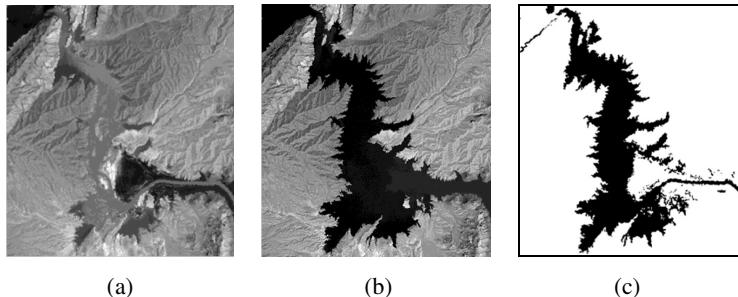


Fig. 4. “Lake Mead”: images acquired in (a) June 1991 and (b) May 2000 (c) ground truth

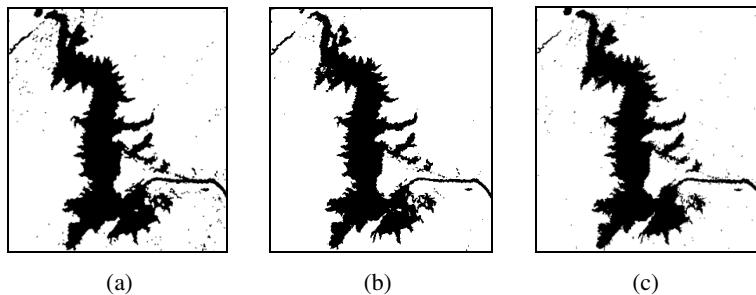


Fig. 5. “Lake Mead”: change detection maps (a) MRF [4] (b) UDWT-AC[3] (c) OMRF

Table 2. “Lake Mead”: comparison between different methods in objective evaluation

Method	Miss Detections		False Alarms		Total Errors	
	Pixels	%	Pixels	%	Pixels	%
MRF[4]	24	0.107	2129	2.402	2153	1.941
UDWT-AC[3]	725	3.251	702	0.799	1427	1.286
OMRF	557	2.497	699	0.788	1256	1.132

5 Conclusion

This paper presents an unsupervised change detection method for multi-temporal remote sensing image. It assumes that all the pixels in same object have an effect on

each other and the pixels between the different objects do not have any spatial interaction. Based on this assumption, we can get better results than traditional methods based on MRF model. Although the algorithm is able to get superior results and reflects strong robustness for noise, it is susceptible to the influence of the region segmentation results. Our future work is to explore ways to reduce the influence.

Acknowledgments. This work was supported by the Research Fund for the Doctoral Program of Higher Education (No. 20126102110041), the Aeronautical Science Foundation of China (No. 2011ZD53049, No. 20125153025), and the Science and Technology on Electro-optic Control Laboratory.

References

1. Celik, T.: Change Detection in Satellite Images Using a Genetic Algorithm Approach. *IEEE Geoscience and Remote Sensing Letters* 7(2), 386–390 (2010)
2. Bovolo, F., Bruzzone, L., Marconcini, M.: A Novel Approach to Unsupervised Change Detection Based on a Semisupervised SVM and a Similarity Measure. *IEEE Trans. Geosci. Remote Sens.* 46(7), 2070–2082 (2008)
3. Celik, T., Kaikuang, M.: Multitemporal image change detection using undecimated discrete wavelet transform and active contours. *IEEE Trans. Geosci. Remote Sens.* 49(2), 706–716 (2011)
4. Bruzzone, L., Prieto, D.F.: Automatic Analysis of the Difference Image for Unsupervised Change Detection. *IEEE Trans. Geosci. Remote Sens.* 38(3), 1171–1182 (2000)
5. Melgani, F., Bazi, Y.: Markovian Fusion Approach to Robust Unsupervised Change Detection in Remotely Sensed Imagery. *IEEE Geoscience and Remote Sensing Letters* 3(4), 457–461 (2006)
6. Li, X.-C., Zhu, S.-A.: A Survey of the Markov Random Field Method for Image Segmentation. *Journal of Image and Graphics* 12(5), 789–798 (2007)
7. Chen, Y., Cao, Z.: An improved MRF-based change detection approach for multitemporal remote sensing imagery. *Signal Processing* 93(1), 163–175 (2013)
8. Redner, R.A., Walker, H.F.: Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review* 26(2), 195–239 (1984)
9. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(5), 603–619 (2002)
10. Besag, J.: On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. B* 48(3), 259–302 (1986)

A Probability-Based Object Tracking Method

Xu Song^{1,2}, Guoqiang Li², Ying Li^{1,2}, and Yanning Zhang^{1,2}

¹ Shaanxi Province Key Lab. of Speech & Image Information Processing,
Northwestern Polytechnical University, Xi'an 710129, China

² Science and Technology on Electro-optic Control Laboratory, Louyang 471009, China
{songhzwg, gouqli}@163.com, {lybyp, ynzhang}@nwpu.edu.cn

Abstract. Here we take advantage of the signal recovery power of Compressive Sensing (CS) to significantly reduce the computational complexity brought by the high-dimension image data, then an effective and efficient low-dimensional subspace representation of the object is computing by applying Principal Component Analysis (PCA) to a collection of object observations which are low-dimensional vectors derived from CS. An incremental PCA algorithm is used to update this subspace model for characterizing the object appearance changes. Meanwhile, two distances derived from Probabilistic Principal Component Analysis (PPCA): distance from feature space (DFFS) and distance in feature space (DIFS), are used to describe visual similarity between the learned subspace representation model and candidate targets. Comparing with the traditional used reconstruction error, the sum of two distances: DFFS + DIFS, is more accurate and more robust to noises and partial occlusions. Numerous experiment demonstrate that subspace representation model can handle the situation that target objects experience pose changes, scale changes, significant illumination variation, partial occlusions and so on.

Keywords: object tracking, PPCA, subspace representation model, particle filter, dynamic model, observation model, Compressive Sensing (CS).

1 Introduction

Object tracking is one of fundamental problems in the field of computer vision. It prevails in diverse applications such as intelligent surveillance, human-computer interface and vehicle navigation. To develop a robust online tracker, it remains as a tough field of study as the visual appearance of target objects may undergo large variations due to many factors such as illumination changes, pose changes, deformations and occlusions. As a result, effectively modeling the appearance changes of the tracked objects plays an important role in visual tracking.

Ross et al [1] proposed a low-dimensional subspace representation and use an efficient incremental PCA algorithm continually updates that model effectively and quickly, it has achieved relatively good performance, while its application is still limited due to the intensive computation caused by PCA. Hence, in this paper we take advantage of the signal recovery power of Compressive Sensing (CS) to significantly reduce the

computational complexity brought by the high-dimension image data, and then an effective and efficient low-dimensional subspace representation of objects is given by applying Principal Component Analysis (PCA) [3] to a collection of object observations which are low-dimensional vectors derived from CS. We make use of the incremental PCA algorithm proposed by Ross et al [1] to effectively update our subspace representation model. We use two distance: DFFS and DIFS, derived from PPCA [4] and followed definition in the paper [5], to describe the visual similarity between our object subspace representation model and the candidate target. The proposed object subspace model and the incremental PCA algorithm model make the tracking process running at real-time, and the two distances ensure its accuracy comparing with state-of-the-art methods. Our method is led by the bayesian inference framework in which a particle filter is used to propagate sample distributions over time.

2 Particle Filter

Let x_t denote the state variable representing the affine motion parameters of the target at time t . Given a set of observations $y_{1:t} = \{y_1, y_2, \dots, y_t\}$, we calculate the posterior probability $p(y_t | x_t)$ using Bayes' theorem [6]:

$$p(x_t | y_{1:t}) \propto p(y_t | x_t) \int p(x_t | x_{t-1}) p(x_{t-1} | y_{1:t-1}) dx_{t-1} \quad (1)$$

where $p(y_t | x_t)$ is the observation model depicting the likelihood of observing y_t at x_t state, and the dynamic (motion) model $p(x_t | x_{t-1})$ gives the definition of motion randomness between two consecutive states. In particle filter, the posterior $p(x_t | y_{1:t})$ is approximated by a finite set of N weighted particles $\{x_t^i, w_t^i\}_{i=1}^N$,

$$p(x_t | y_{1:t}) = \sum_{i=1}^N w_t^i \delta(x - x_t^i) \quad (2)$$

We give the detail of particle filter used in our tracking formwork below.

2.1 Dynamic Model

In this paper, we apply an affine image warp to model the target state x_t with six parameters. Let $x_t = \{x_t, y_t, \sigma_t, sx_t, sy_t, r_t\}$ be affine warp parameters. The dynamic model is formulated by a Gaussian distribution as follow:

$$p(x_t | x_{t-1}) = N(x_t; x_{t-1}, \Phi) \quad (3)$$

where Φ is a diagonal covariance matrix whose diagonal elements depict that how much we expect the target object might move from one frame to the next. Actually, the value of Φ 's diagonal elements affect the accuracy of our tracking process, with smaller values in the diagonal covariance matrix Φ , we may lost the target, and larger values may need more particles.

2.2 Observation Model

We crop the region of one candidate particle x_t^i from the current image, and normalize it to be the standard size $M * N$, y_t^i is a certain kind of feature extracted from the cropped region. Then the probability of y_t^i being generated from our object representation model is depicted by the observation model $p(y_t | x_t)$. We give the definition of $p(y_t | x_t)$ based on PPCA and the structure of our object representation model below.

3 PPCA and Initial Form of Our Subspace Model

The tracking problem can be described as follows: given t observations $\{y_i\}_{i=1}^t$ (the tracked targets in the first t frame images), and N particles $\{y_{t+1}^i\}_{i=1}^N$, we want to calculate the probabilities of $\{y_{t+1}^i\}_{i=1}^N$ being generated from that target class. Assumed that $\{y_i\}_{i=1}^t$ obey the Gaussian distribution, then compute the mean and covariance matrix, thusly, the problem is solved. As the observation y is often a very high-dimensional feature vector, it's time-consuming to calculate the covariance matrix. Thusly, we turn to PPCA which uses a low-dimensional latent variable model to represent the high-dimensional feature space. At the same time we get an initial form of our subspace representation model.

3.1 A Latent Variable Model

A latent variable model seeks to relate a r -dimensional observation vector y to a corresponding d -dimensional vector of latent variables z ($r < d$). The most common used model is factor analysis where the relationship is followed:

$$y = Wz + u + \varepsilon \quad (4)$$

where $W \in R^{r*d}$ is projection matrix, u is the mean of y , ε is an additional noise. As commonly assumed in factor analysis: $z \sim N(0, I_d)$, $\varepsilon \sim N(0, \sigma^2 I_r)$, we can calculate the prior probability of the observation vector y from Eq.4,

$$p(y) \sim N(y; u, WW^T + \sigma^2 I_r) \quad (5)$$

Thusly, a low-dimensional latent vector is used to represent a high-dimensional data space in a probabilistic manner. Eq.5 is the corresponding explanation in PPCA of the observation model $p(y_t | x_t)$ of object tracking formwork. So we have:

$$p(y_t | x_t) = N(y_t; u, WW^T + \sigma^2 I_r) \quad (6)$$

Assumed that $Y = \{y_i\}_{i=1}^t$ are training sample set collected from the past t object image patches, its covariance matrix is C , Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ are r eigenvalues of C , $\Sigma_d = diag(\lambda_1, \lambda_2, \dots, \lambda_d)$ is the diagonal matrix composed by the d largest

eigenvalues, U_d is the corresponding eigenvectors of Σ_d . Tipping and Bishop [4] give the MLE of u , W , σ^2 :

$$u = \frac{1}{t} \sum_{i=1}^t y_i, W = U_d (\Sigma_d - \sigma^2 I_d)^{1/2} R, \sigma^2 = \frac{1}{r-d} \sum_{i=d+1}^r \lambda_i \quad (7)$$

where $R \in \mathbb{R}^{d \times d}$ is an arbitrary orthogonal rotation matrix. We should note here that U_d , Σ_d and u form our subspace model which is spanned by U_d and centered at u , and y is the observation compressed by sparse measurement matrix P .

3.2 Probability Computation with PPCA

When Eqs.7 are known, we can compute the probability Eq.6. By taking the log of Eq.6 and removing the constant part, we have the log-probability followed:

$$L = (y - u)^T U_d \Sigma_d^{-1} U_d^T (y - u) + \frac{1}{\sigma^2} (y - u)^T (I_r - U_d U_d^T) (y - u) \quad (8)$$

The conditional probability $p(y_t | x_t)$ is decided by the two parts of Eqs.8. The first part is the Mahalanobis distance [6] of y spanned by U_d , and the second part is the vertical distance from y to the subspace spanned by U_d , centered at u . We refer to the two distances as DIFS and DFFS which are followed the definition of paper [5].

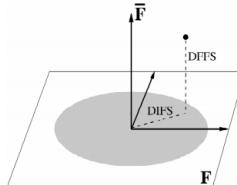


Fig. 1. DFFS and DIFS, F is the hyper-plane, \bar{F} is its orthogonal complement, the black point is the location of y in feature space, and the gray region in the hyper-plane is the subspace constructed by U_d , Σ_d and u

From Fig.1, we can see that DIFS and DFFS are more accurate than only using reconstruction error (Actually it's DFFS) of PCA. That is why we use DIFS and DFFS to calculate the conditional probability $p(y_t | x_t)$.

4 Online Updating Our Subspace Representation Model

In the previous section, we take U_d , Σ_d and u as our object subspace representation model. During the tracking, the training samples arrived one by one, so we need an online updating algorithm to update U_d , Σ_d and u . We use the incremental PCA algorithm proposed in paper [1], here is the main steps of the algorithm:

The incremental PCA algorithm

Assumed: $A = [y_1, y_2, \dots, y_n]$ is a $r \times n$ matrix where each column y_i is the observation extracted from the tracked target region in the i th frame. A $r \times m$ matrix $B = [y_{n+1}, y_{n+2}, \dots, y_{n+m}]$ is the new tracked targets' observations collected from the new tracked frames. Let $C = [A, B]$ represent the combination of A and B . Denote the mean of A , B , C as I_A , I_B , I_C . U_d , Σ_d and I_A construct of our subspace model are known.

Input: B and I_B , U_d , Σ_d computed from the SVD of $A - I_A$.

Output: new subspace model based on C : U' , Σ' , I_C .

1. Compute $I_C = \frac{f \cdot n}{f \cdot n + m} I_A + \frac{m}{f \cdot n + m} I_B$, f is forgetting factor, $0 < f \leq 1$;
 2. Let $\bar{B} = [I_{m+1} - I_B, \dots, I_{m+n}, \sqrt{\frac{n \cdot m}{m+n}}(I_B - I_A)]$;
 3. Obtain $[B^*, V] = QR(\bar{B} - U_d U_d^T \bar{B})$, $QR(\bullet)$ is a QR decomposition process;
 4. Let $R = \begin{bmatrix} f\Sigma_d & U_d^T \bar{B} \\ 0 & B^{*T}(\bar{B} - U_d U_d^T \bar{B}) \end{bmatrix}$, compute the SVD of R : $R = \bar{U} \bar{\Sigma} \bar{V}^T$;
 5. Finally, $U' = [U_d, B^*] \bar{U}$, $\Sigma' = \bar{\Sigma}$.
-

5 Compressive Sensing for Dimensionality Reduction

Although the incremental PCA algorithm can efficiently update the subspace representation model, we find the potential acceleration power though analyzing the process of incremental PCA algorithm.

Here we use a sparse measurement matrix that satisfies the Restricted Isometry Property (RIP) [7] to project the image feature space to a low-dimensional compressed subspace. And the Johnson-Lindenstrauss lemma [8] states that with high probability the distances between the points in original feature space are preserved if they are projected onto that low-dimensional compressed subspace.

The sparse measurement matrix $P \in R^{L \times H}$ ($L \ll H$) used is defined as follow [9].

$$p_{i,j} = \begin{cases} 1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ -1 & \text{with probability } \frac{1}{2s} \end{cases} \quad (9)$$

$s=2$ or $s=3$ and $i=1,2,\dots,L$, $j=1,2,\dots,H$

Assumed that v is high-dimensional feature vector extracted from target object image region (in our tracking algorithm, v is the result of original target object image region convolved by different scale rectangle filter), and the observation

$$y = Pv \quad (10)$$

At last, we give the final form of subspace representation model which is constructed of K largest eigenvalues centered by the mean and its corresponding eigenvalues computed from the compressed observation y . We can still use the incremental PCA to update that subspace model and calculate visual similarity between that subspace model and candidate targets as being mentioned earlier, while ours method is more faster.

6 Experimental Results

To demonstrate the performance of our tracking algorithm, we tested it on 6 image sequences (car4, David, deer, lemming, occlusion, girl) collected from the public dataset. Comparing with 3 classic tracking methods (Frag-Track [10], CT [9], l_1 tracker [11]), we demonstrate the superiority of our tracking algorithm.

In the Lemming sequence, our algorithm, Frag-Track, CT and l_1 tracker work well in the first few frames, but when the target (a toy) undergoes fast motion and blurred, Frag-Track, CT and l_1 tracker lost target, and our algorithm tracks the target accurately. Also ours is able to handle occlusion better. As be shown in figure 2.

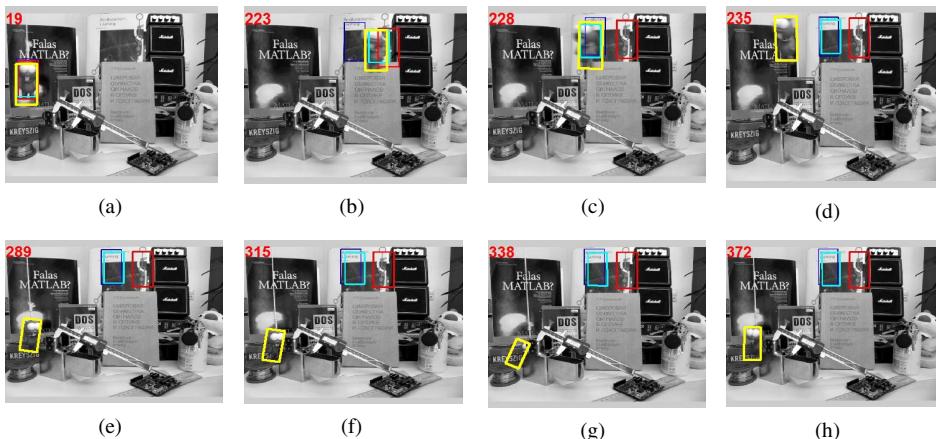


Fig. 2. The Lemming sequence. Red solid line – Frag-Track, blue solid line – CT, green solid line – l_1 tracker, yellow solid line –Ours.

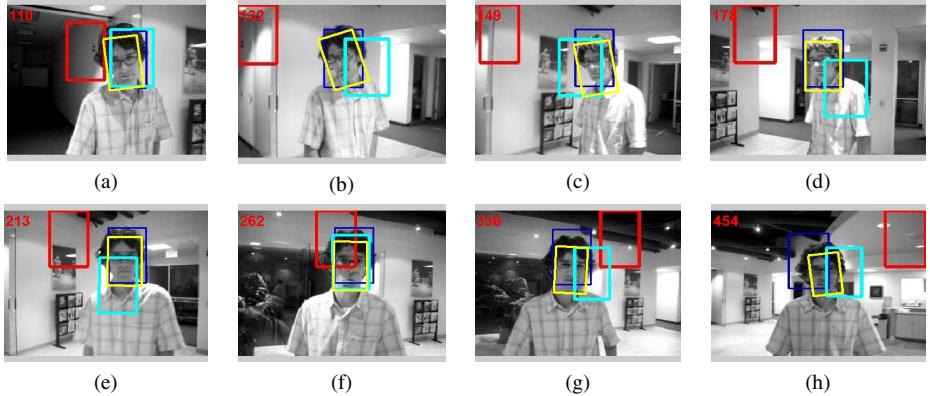


Fig. 3. The David sequence. Red solid line – Frag-Track, blue solid line – CT, green solid line – l_1 tracker, yellow solid line – Ours.

Figure 3 gives the results of four tracking algorithm in the David sequence which target object undergoes illumination changes, rotation, appearance variations and scale changes. Frag-Track, l_1 tracker gradually lose the target from 110th-frame where large appearance changes and rotations occurred. CT tracker and our algorithm can track long time, but CT tracker showed drift problem.

We use Average Center Error (ACE) for performance evaluation. And we present the results in Table.1. The results show that our algorithm is only suboptimal in Car sequence and optimal averagely.

Table 1. Average Center Error (Pixels). The best results are shown in bold and green font.

	Frag-Track	CT	l_1 tracker	Ours
Lemming	149.1	135.9	184.9	7.2
David	76.7	5.9	7.6	3.9
Car	179.8	153.7	4.1	4.5
Deer	92.1	64.8	71.5	12.6
Girl	48.7	20.6	24.8	5.4
Occlusion	5.6	4.3	6.5	3.4
Average	92	64.2	49.9	6.17

7 Conclusion

In this paper, we propose a very simple object subspace representation model to deal with the high-dimensional image feature space. The subspace model achieves the real-time updating speed. And we use two distance: DFFS and DIFS, derived from PPCA to depict visual similarity between target objects and our subspace model that is more accurate when we apply the maximum a posteriori (MAP) estimation to object state

estimation. Numerous experiment demonstrate that subspace representation model can handle the situation that target objects experience the long time appearance variations caused by many intrinsic and extrinsic reasons, such as pose changes, scale changes, significant illumination variation, partial occlusions and so on.

Acknowledgement. This work was supported by the Research Fund for the Doctoral Program of Higher Education (No. 20126102110041), the Aeronautical Science Foundation of China (No. 2011ZD53049, No. 20125153025), and the Science and Technology on Electro-optic Control Laboratory.

References

1. Ross, D., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. *IJCV* 77(13), 125–141 (2008)
2. Donoho, D.: Compressed sensing. *IEEE Trans. Inform. Theory* 52, 1289–1306 (2006)
3. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*. Academic, New York (1990)
4. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* 61(3), 611–622 (1999)
5. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19(7) (1997)
6. http://en.wikipedia.org/wiki/Mahalanobis_distance
7. Candes, E., Tao, T.: Decoding by liner programing. *IEEE Trans. Inform. Theory* 51, 4203–4215 (2005)
8. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.* 66, 671–687 (2003)
9. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part III. LNCS*, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
10. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 798–805 (2006)
11. Mei, X., Ling, H.: Robust visual tracking using L1 minimization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1436–1443 (2009)

Pedestrian Detection Based on Incremental Learning

Yu Xia¹, Yongzhen Huang², Liang Wang², and Xin Geng^{1,*}

¹ School of Computer Science and Engineering, Southeast University, Nanjing, China
xgeng@seu.edu.cn

² National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China

Abstract. Pedestrian detection is a hot topic in computer vision and pattern recognition. Existing pedestrian detection methods face new challenges in the background of big data, e.g., heavy burdens on computing and memory. To solve these problems, in this paper, we propose a pedestrian detection framework based on incremental learning. Compared with existing pedestrian detection frameworks, it costs much less time and memory. In addition, the performance of our framework is very close to the one which uses all training samples at once. Furthermore, with more new training samples, the performance can be enhanced continually with little time and memory, showing the potential in practical applications.

Keywords: Pedestrian detection, Incremental learning, Converged passive-aggressive, Histograms of oriented gradients.

1 Introduction

Pedestrian detection is an important issue in computer vision and can be applied in many related areas [4,5]. Many effective algorithms are developed. For example, Viola *et al.* [13,14] propose object detection algorithms based on Haar-like features and the cascade structure. However, Haar-like descriptors are not able to describe more complicated objects (such as pedestrians under complex backgrounds). To obtain more robust descriptor, Dalal *et al.* [3] propose HOG (histograms of oriented gradients) for pedestrian detection. HOG describes shape and structure well, and obtains satisfying results in pedestrian detection. However, this approach takes no account of local deformation of objects. To solve this problem, Felzenszwalb *et al.* [6] propose a discriminative part-based approach that models unknown part positions as latent variables. It is one of the most successful approaches for object detection at present.

In the background of big data, it is possible to collect huge pedestrian data sets, which are usually captured continually. Existing pedestrian detection methods do not perform well in this situation and face some new challenges.

* Corresponding author.

Think about this case: collecting new training sets from various sources (e.g., web, cameras, videos and so on) continuously after training a detector using former training sets, and updating the detector with new training sets. Traditionally, we need to merge the former and the new training sets, and use the merged training set to train a new detector. In this case, existing methods consume huge time when data sets are very large, e.g., million or billion. What's more, limited memory is not enough to train such a detector. The amount of data is always too large for machines to handle. It is necessary to study incremental learning based pedestrian detection which makes feasible in case of big data, runs fast and maintains satisfying performance.

There is little work on embedding incremental learning into general object detection. Opelt *et al.* [10] introduce incremental learning into visual shape alphabet based multi-class object detection. This method is designed for a small or medium size of training sets but not suitable for big data. Nair *et al.* [7,8,9] combine online learning and object detection in video with simplex and still background. These methods are not effective for general object detection with complex background.

In this paper, we propose an incremental algorithm which is simple but effective. We apply it to pedestrian detection and obtain satisfying results. Our framework has three merits: (1) running very fast; (2) needing only a small size of memory; (3) improving the detection accuracy continually. Our algorithm achieves good performance on the INRIA and the NICTA pedestrian datasets [3,11] in terms of both effectiveness and efficiency, which demonstrates that our algorithm is suitable for pedestrian detection with huge data.

2 Pedestrian Detection Based on CPA

2.1 Framework

The HOG based pedestrian detection algorithm [3] is probably the first pedestrian detection approach which can achieve satisfying performance for real images with complex backgrounds [4,5]. In Fig.1, the diagram inside the dotted line box shows the procedure of HOG based pedestrian detection. Although this approach is less accurate than the part-based model [6], its complexity is far lower and thus it is very suitable to be combined with incremental learning. This is the main reason why we adopt HOG based pedestrian detection as our baseline platform.

Fig.1 shows the framework of incremental learning based pedestrian detection. Intuitively, this framework has two advantages. On one hand, if the data sets are collected continuously, the system could rapidly and effectively train detectors using new data sets. On the other hand, if a data set is so large that memory can not meet the demand, we could divide it into several parts and use them to train a detector according to the procedure in Fig.1.

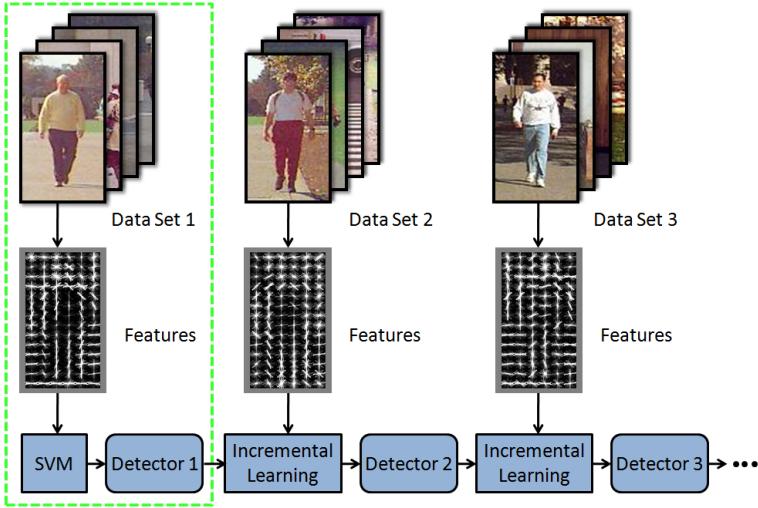


Fig. 1. Framework of incremental learning based pedestrian detection

2.2 Incremental Learning Algorithm

Compared with batch learning algorithms, incremental learning algorithms run faster and cost much less memory. We propose a new incremental learning algorithm called Converged Passive-Aggressive algorithm (CPA) for pedestrian detection.

CPA is based on the Passive-Aggressive algorithm [2]. The objective function of Passive-Aggressive algorithm is,

$$\mathbf{W}_{t+1} = \operatorname{argmin}_{\mathbf{W} \in \mathbb{R}^n} \frac{1}{n} \|\mathbf{W} - \mathbf{W}_t\|^2 + C\xi^2 \quad \text{s.t.} \quad \ell(\mathbf{W}; (\mathbf{X}_t, y_t)) \leq \xi. \quad (1)$$

Here \mathbf{W}_t is the vector containing parameters of the detector, \mathbf{X}_t is the sample, and y_t is the label of the sample in round t . ℓ is hingle-loss function, ξ is a non-negative slack variable, and C is the weight of ξ^2 . \mathbf{W}_{t+1} is the projection of \mathbf{W}_t onto the half-space of vectors whose hingle-loss is zero for the current sample.

If $\ell(\mathbf{W}; (\mathbf{X}_t, y_t)) > 0$, the Lagrangian of the problem in (1) is

$$L(\mathbf{W}, \xi, \alpha) = \frac{1}{2} \|\mathbf{W} - \mathbf{W}_t\|^2 + C\xi^2 + \alpha(1 - \xi - y_t(\mathbf{W}^T \mathbf{X}_t)). \quad (2)$$

Here α is a non-negative Lagrange multiplier. Setting the partial derivatives of L with respect to \mathbf{W} and ξ to zero respectively, we obtain

$$\mathbf{W} = \mathbf{W}_t + \alpha y_t \mathbf{X}_t. \quad (3)$$

$$\xi = \frac{\alpha}{2C}. \quad (4)$$

Replacing ξ and \mathbf{W} with (3) and (4), the Lagrangian can be expressed as

$$L(\alpha) = -\frac{\alpha^2}{2}(\|\mathbf{X}_t\|^2 + \frac{1}{2C}) + \alpha(1 - y_t(\mathbf{W}_t^T \mathbf{X}_t)). \quad (5)$$

Setting the derivative of $L(\alpha)$ to zero, we obtain

$$\alpha = \frac{1 - y_t(\mathbf{W}_t^T \mathbf{X}_t)}{\|\mathbf{X}_t\|^2 + \frac{1}{2C}}. \quad (6)$$

It should be noted that each sample is used only once so that the solution may not be convergent. To address this problem, a weight-decay strategy is proposed as shown in Algorithm 1.

In Algorithm 1, C_{INIT} is an initial parameter. dr is a scalar which regulates the decreasing rate of C . M is the number of samples. K is the number of iterations in the outer loop. Crammer *et al.* [2] show that setting C to be a small number leads to a slow progress rate and gets better solution when training sets are large. So, C is set to decrease in each round. Generally, the outer loop of CPA needs a small number of (less than 20) iterations.

Algorithm 1. CPA

Input:

Initial Passive-Aggressive parameter, C_{INIT} ;

Decreasing rate of C_{INIT} , parameter dr ;

Initial model $\mathbf{W}_{1,1}$;

Output:

Updated model $\mathbf{W}_{K,M+1}$.

for $k = 1, 2, \dots, K$ **do**

$C = \frac{C_{INIT}}{dr \times k}$

for $t = 1, 2, \dots, M$ **do**

receive instance: $\mathbf{X}_t \in \mathbf{R}^n$

receive correct label: $y_t \in \{-1, +1\}$

compute loss: $\ell_t = \max\{0, 1 - y_t(\mathbf{W}_{k,t}^T \mathbf{X}_t)\}$

compute Lagrange multiplier: $\alpha_{k,t} = \frac{\ell_t}{\|\mathbf{X}_t\|^2 + \frac{1}{2C}}$

update: $\mathbf{W}_{k,t+1} = \mathbf{W}_{k,t} + \alpha_{k,t} y_t \mathbf{X}_t$

end for

$\mathbf{W}_{k+1,1} = \mathbf{W}_{k,M+1}$

end for

3 Experiments and Results

3.1 Experimental Datasets and Setup

Experiments are conducted on two different pedestrian datasets. The first is the well known INRIA pedestrian dataset [3], containing 3,542 (2,416 for training

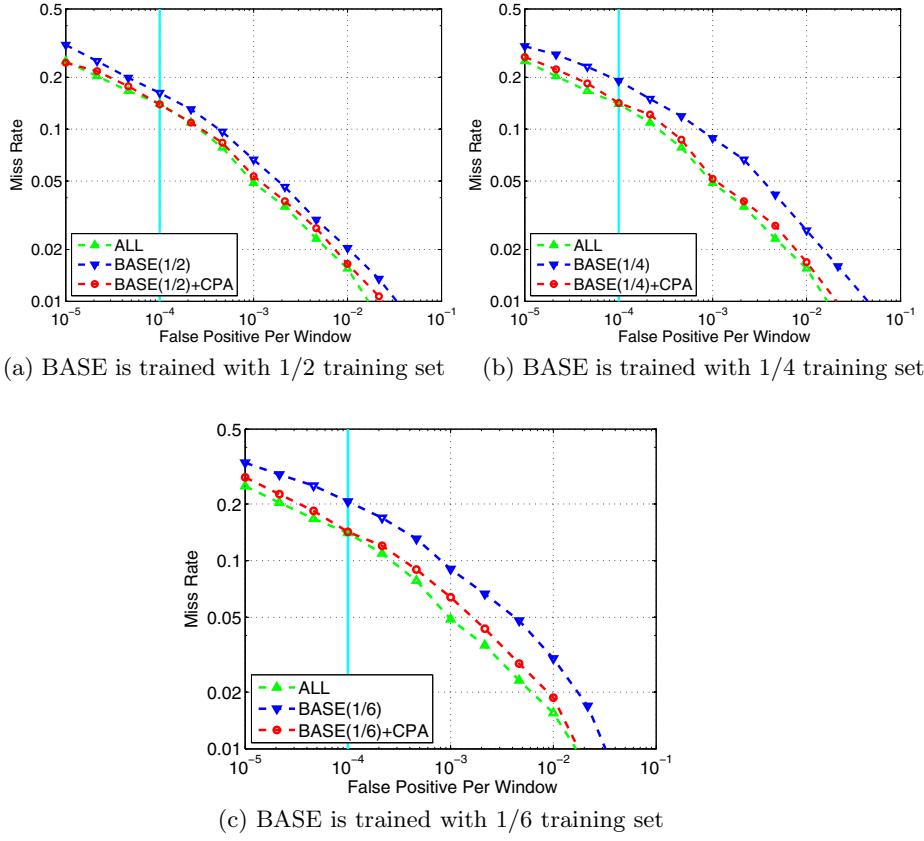


Fig. 2. Performance of CPA with different initial detectors

and 1,126 for testing) positive samples and 1,671 (1,218 for training and 453 for testing) background images. The resolution of pedestrian images is 128×64 pixels. We use this database to evaluate the major characteristics of our proposed algorithm. The second is the NICTA pedestrian dataset [11], containing 44,223 (37,344 for training and 6,879 for testing) positive samples and 5,151 (4,147 for training and 1,004 for testing) background images. The resolution of pedestrian images is 80×32 pixels. We use this database to simulate the performance of our framework in the case that training samples are continually obtained.

Because the resolution of images in the two datasets is different, we set the size of each cell to 8×8 pixels on the INRIA dataset and 4×4 pixels on the NICTA dataset. To initialize, we use a soft linear SVM trained with libSVM [1].

3.2 Results on INRIA

Experiments are designed to study the learning ability of CPA in various situations. First, dividing the training set into two parts evenly, we use the first part

to train a linear SVM and obtain an initial detector denoted as BASE. Then we use the other part to update the detector with CPA and obtain a detector denoted as BASE+CPA. To compare the performance of CPA, all training samples of the INRIA database are also used to train a linear SVM, and get a detector denoted as ALL. Fig.2(a) shows the DET (Detection Error Tradeoff) curves [3] of BASE, BASE+CPA and ALL.

Table 1. Time(sec) consumed by each detector

	BASE(1/2)	BASE(1/4)	BASE(1/6)	ALL
BASE	56.05	15.60	8.75	145.74
BASE+CPA	56.05+1.57	15.60+1.23	8.75+1.37	-

Then, we do experiments to compare the performance of BASE+CPA with different initial detectors. Dividing the training set into four parts evenly, we use the first part to train a linear SVM and obtain an initial detector. Then we use the rest three parts to update the detector with CPA three times. Each time we use one part. Fig.2(b) shows the DET curves of the detectors. Likewise, dividing the training set into six parts evenly, we train the detector as the previous procedures. This result is shown in Fig.2(c). In addition, the time costed by training each detector is shown in Table 1. From Fig.2 and Table 1, it is easy to obtain the following conclusions:

1. The more training samples, the better the detector performs. For example, the results is gradually enhanced over detectors BASE(1/6), BASE(1/4), BASE(1/2) and ALL.
2. No matter which baseline detector is adopted, our proposed CPA can enhance the baseline effectively and finally performs comparably with the detector ALL.
3. CPA costs very little time compared with SVM, which means that CPA is very efficient to reach satisfying detection accuracy.

We also compare CPA with two incremental learning algorithms (Pegasos [12] and Passive-Aggressive algorithm) on the INRIA database. Pegasos is a popular incremental learning algorithm for linear SVMs. Table 2 shows the miss rate at 10^{-4} FPPW (false positive per window) of different detectors. The results of CPA are the best in all cases. On average, CPA has above 1.5% lower miss rate than Passive-Aggressive and above 0.8% lower miss rate than Pegasos.

3.3 Results on NICTA

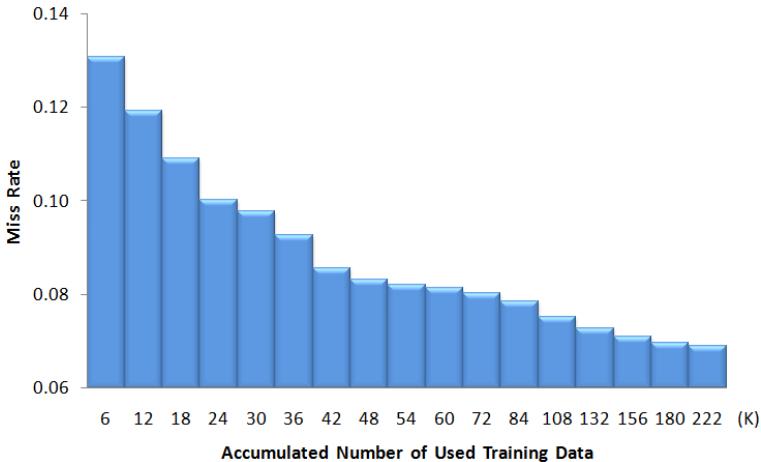
To simulate incremental learning based pedestrian detection with training data continually captured, experiments are designed on the NICTA pedestrian database. First, using 1,000 positive and 5,000 negative samples to train a linear SVM, we obtain an initial detector. Then, we use new data (the ratio of positive

Table 2. Miss rate of different detectors at 10^{-4} FPPW

	BASE(1/2)	BASE(1/4)	BASE(1/6)
Passive-Aggressive	15.63%	15.90%	15.36%
Pegasos [12]	14.65%	15.01%	15.19%
CPA	13.94%	14.12%	14.21%

samples to negative samples is 1/5) to train the detector continually as the procedures in Fig.1.

Fig.3 shows that the miss rate at 10^{-4} FPPW decreases with incrementally learning from new data. After initializing a detector, we use new training data to update the detector 16 times. Miss rate declines continually, which demonstrates that our framework has good ability for training better detectors.

**Fig. 3.** The performance of the detector with increasing new training samples

4 Conclusions

In this paper, we have proposed a pedestrian detection framework with incremental learning. The detection accuracy of our method is comparable to traditional pedestrian detection which uses all training samples at once, but consumes much less time and memory. Furthermore, we demonstrate that the detector can be continually enhanced with more new training samples, which is very valuable for practical applications.

Finally, we conclude this paper with two main contributions. First, we have proposed a new incremental learning algorithm CPA to enhance online Passive-Aggressive algorithm. Second, we have embedded the enhanced incremental learning algorithm into pedestrian detection to effectively and efficiently solve problems of existing pedestrian detection methods in the background of big data.

Acknowledgment. This research was supported by the National Science Foundation of China (61273300, 61232007, 61203252, 61175003), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Excellent Young Teachers Program of Southeast University, the Key Lab of Computer Network and Information Integration of Ministry of Education of China.

References

1. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(27), 1–27 (2011)
2. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. *The Journal of Machine Learning Research* 7, 551–585 (2006)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893 (2005)
4. 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)
5. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(12), 2179–2195 (2009)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1627–1645 (2010)
7. Grabner, H., Bischof, H.: On-line boosting and vision. In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 260–267 (2006)
8. Javed, O., Ali, S., Shah, M.: Online detection and classification of moving objects using progressively improving detectors. In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 696–701 (2005)
9. Nair, V., Clark, J.J.: An unsupervised, online learning framework for moving object detection. In: *Computer Vision and Pattern Recognition*, vol. 2, pp. 310–317 (2004)
10. Opelt, A., Pinz, A., Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet. In: *Computer Vision and Pattern Recognition*, vol. 1, pp. 3–10 (2006)
11. Overett, G., Petersson, L., Brewer, N., Andersson, L., Pettersson, N.: A new pedestrian dataset for supervised learning. In: *IEEE Intelligent Vehicles Symposium*, pp. 373–378 (2008)
12. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. In: *International Conference on Machine Learning*, pp. 807–814 (2007)
13. Viola, P., Jones, M.J.: Robust real-time face detection. *International Journal of Computer Vision* 57(2), 137–154 (2004)
14. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision* 63(2), 153–161 (2005)

Vanishing Point Detection Based on Infrared Road Images for Night Vision Navigation

Huan Wang, Feifei Li, and Mingwu Ren

Nanjing University of Science and Technology, Nanjing, P.R. China
wanghuanphd@foxmail.com, feifeili@gmail.com, renminwu@sina.com

Abstract. Road detection is important in computer vision for autonomous driving, pedestrian detection and other applications. Visible light (VL) camera is often used for daytime road detection, and infrared camera is often used for night road detection. Vanishing point (VP) detection is useful for inferring road region. Many VP detection methods have been proposed and applied successfully in VL road image. However, there is no special VP detection method for infrared road image. In this paper, we propose a VP detection approach for infrared road detection. The novelty of our approach relies on the rational assumption that the regions are very similar along the direction of the true VP. This assumption is involved in finding effective VP voters by using a non-local similarity manner, and these VP voters estimate the VP together. Quantitative and qualitative experiments show the effectiveness and efficiency of the proposed method.

Keywords: Vanishing point detection, Infrared Road detection, Non-local similarity.

1 Introduction

Road detection is an important research topic in the area of transportation systems, such as autonomous driving [1-2], driver assistance systems [3-4] and so on. Visible light (VL) camera is often used for daytime road detection [5], and infrared camera is often used for night road detection [6]. Vanishing point (VP) detection is useful for inferring road region. Many VP detection methods have been proposed and applied successfully in VL road image [7-8]. The state of the art vanishing point detection method is based on texture analysis, such as Gabor filter[9], Laplacian of Gaussian Filters[10] or edge detection [11]. However, there is no special VP detection method for infrared road image. In addition, VP detection based on texture analysis is hard for infrared images since infrared image usually has weak texture, and the edge magnitude is also small since the temperature of road and non-road region varies smoothly.

In fact, we observe that the road boundary or some of objects in a road scene usually have high similarity along the direction of VP. For a road region in an image, we can find regions similar with the region in all direction. For road boundary

regions, we can only find high similarity along the direction of vanishing point. For some of non-road region, we can not find high similarity regions or we can find similar regions in all direction similar to the road region, for other non-road regions, we can also find similarity regions along the direction of VP similar to boundary regions. Applying this prior, we design an effective and efficient VP detection method for infrared road image. This idea is very similar to the principle of non-local similarity, which has been applied successfully in image de-noising [12] and image representation [13]. Their difference is the aim of finding several similar regions in non-local regions is to investigate their space distribution, that is, the centers of similar regions distribute in a line type or a cluttered type, other than to calculate the average of similar regions or the representation for an image patch.

The primary superiority of our method is that it needs no magnitude threshold selection for edge detection and scale parameter turning for texture analysis, instead, the intensity information of an original image is used directly.

2 Proposed Method

Our approach broadly consists of three stages: high contrast point extraction, voter point selection and vanish point voting. Below, we describe each stage in detail.

2.1 High Contrast Point Extraction

Considering road region is always in the bottom of the road image in vision navigation scene, so the first preprocessing step consists the generation of a mask that restricts the area of analysis, as shown in Fig.1(b), or we call region of interest (ROI).

Beside, computational efficiency is necessary and considered a lot, so we sample pixels where the region centered at this pixel has high contrast as processing objects, generally, these pixels and its surrounding region are very useful for VP voting. Ref.[14] proposed a simple but effective method for high contrast region extraction, it uses a simple measure of local contrast which is defined as the local standard deviation s of the image intensities divided by local mean intensity u :

$$c = \frac{s}{u} \quad (1)$$

where:

$$u = \frac{1}{w \cdot h} \sum_{i=1}^h \sum_{j=1}^w f_{ij} \quad (2)$$

$$s = \sqrt{\frac{1}{w \cdot h} \sum_{i=1}^h \sum_{j=1}^w (f_{ij} - u)^2} \quad (3)$$

w, h represent image width and height, f_{ij} is denoted as intensity of the pixel in the image coordination (i, j) we set $w = h = 11$ in our experiments.

We find that this method is insensitive to image noise. We applied this method for our ROI extraction. For each pixel p_i in the ROI of original infrared image, we calculate c_i according to equation (1). Fig.1(c) show the contrast map of Fig.1(a).

Then we use calculate a threshold T to select high contrast points based on follow equations.

$$H_{p_i} = \begin{cases} 1 & c_i > T \\ 0 & \text{else} \end{cases} \quad (4)$$

$$T = \max(\epsilon, u' + s') \quad (5)$$

where $H_{p_i} = 1$ means p_i is a high contrast point, otherwise not, u' and s' are the mean and the standard deviation of the set $\{c_i\}$ respectively, their calculation is similar to equation (2) and (3), ϵ is a small constant, we set $\epsilon = 3$ in our experiments. Fig.1(d) show extracted high contrast points.

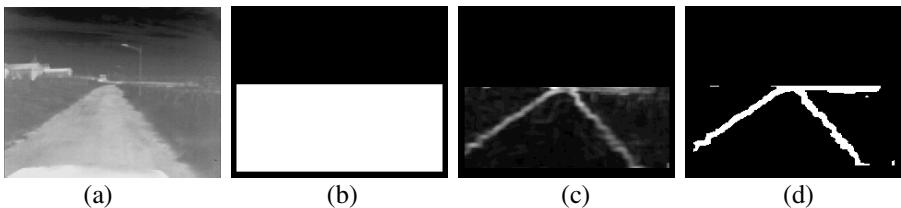


Fig. 1. High contrast points extraction. (a) an infrared road image, (b) mask image depicting region of interest,(c) contrast map, and (c) high contrast points (white points).

2.2 Voter Point Selection

Now, a set of high contrast points are obtained, we further select points which can be used for VP voting from these high contrast points, the new selected points are called voter points.

For road image, the road pavement is always regular along the direction of VP, especially for some off-road regions. We often find that there are sidewalks, water ditch, vegetation, along the two side of the road, they are companion with the road, and they will give good inferring information for VP detection. Due to the perspective projection, these regions are distributed along the direction of VP.

In infrared image, a high contrast point being a voter point should pass following test, that is, if we search in a fixed rectangle region, which is above current high contrast point, as the white rectangles shown in Fig.2(a), we will find several similar regions. The search region and search strategy is importance, if we find similar region only in the neighborhood region of a pixel, it will lead to unbelievable result since region has local similarity. We select the most similar region in each row of the search regions (white rectangle), and we select the top N regions based on their

similarity. Generally, correlation coefficient is often used to measure the similarity of two regions:

$$R(x, y) = \frac{\sum_{y=0}^H \sum_{x=0}^W (M(y, x) - \bar{M})(I(y, x) - \bar{I})}{\sqrt{\sum_{y=0}^H \sum_{x=0}^W (M(y, x) - \bar{M})^2} \sqrt{\sum_{y=0}^H \sum_{x=0}^W (I(y, x) - \bar{I})^2}} \quad (6)$$

Where M and I are two image patch, \bar{M} and \bar{I} are their mean respectively.

In fact, we find the absolute difference (equation (7) also works and it is more efficient than correlation coefficient.

$$dif = \sum_{y=0}^H \sum_{x=0}^W |M(y, x) - I(y, x)| \quad (7)$$

So we applied absolute difference to measure similarity. If the dif is small than threshold (we set 3 in our experiments), we consider two comparison regions are similar. After obtaining similar regions, we further judge the position relationship of center points of these regions. First, these regions should aggregate in a line type, second, the number of these regions should large than half of the selected similar regions. If the two conditions are satisfied, a voter point is found. Fig.2(a) show two regions (red rectangles) center at their voter points and their similar regions (blue rectangles).

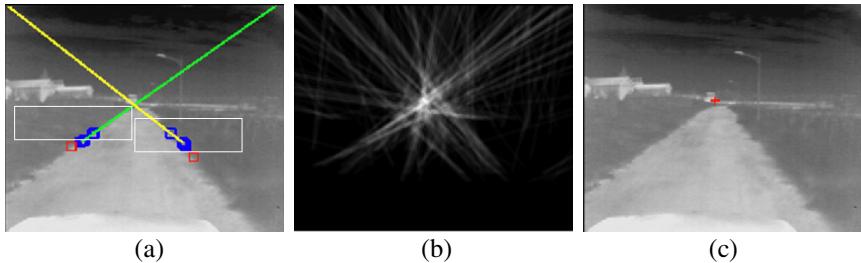


Fig. 2. Vanishing point detection. (a) two regions center at their voter (red rectangles) point and their similar regions(blue rectangles), and the fitting line using center point of these similar regions (yellow and green line), (b) voting accumulation map, and (c) detected vanishing points (red points).

2.3 VP Voting

For each voter point and the center points of its non-local similar regions, we can fit a straight line use minimum distance error criteria.

$$e = \min \frac{1}{N} \sum_{\substack{(x, y) \in \\ (x_i^j, y_i^j)_{j=1}^N}} \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}} \quad (8)$$

That is, for any two points, a straight line is fit. Then we calculate the distance error from a point (x, y) to the fit straight line equation $Ax + By + C = 0$, and average distance is obtained. The straight line which takes the minimum average distance error e is selected.

By generating an accumulation image initialized with zero, all the points in the fit straight line are accumulated in the accumulation image. The accumulation image of Fig.2(a) is shown in Fig.2(b), and then the position with maximum accumulation value in the accumulated image is considered as the VP, as shown in Fig.2(c),

3 Experiments

All the experimental data in this paper are captured by a infrared camera which is mounted in the front floor of our autonomous lane vehicle. The image resolution is 352 by 288. Some typical detection results are reported in Fig.1, where the six figures depict different scenes include two type pavement, sandstone road and cement road, and different non-road objects, such as trees, vegetations, sidewalk involved with shadows and so on. Red crossing show their VP detection results.

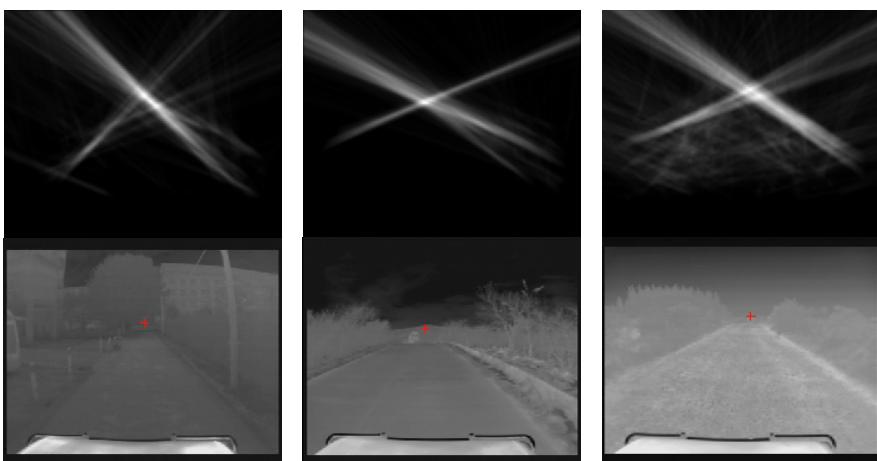


Fig. 3. Results of vanishing point detection in different road scene

We also select two state of the art VP detection methods for comparison. The two methods are proposed in Ref.[9] and Ref.[10]. One is based on Gabor filter, and the other is based on Laplacian of Gaussian Filters. It should be noted here that the two methods for comparison obtain similar results as our method for the images shown in Fig.3. However, they are not robust for other images in our database. We show some VP detection results which the two methods fail but our method succeeds in Fig.4.

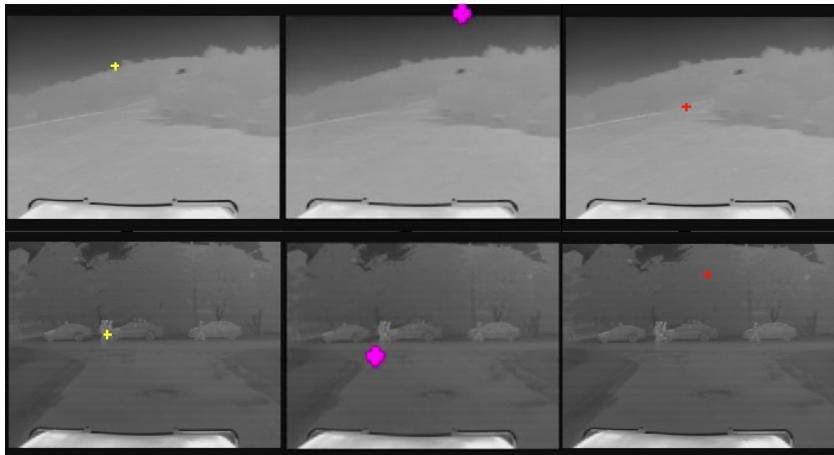


Fig. 4. Comparison of the three methods. (a) method proposed in Ref.[9], (b) method proposed in Ref.[10], (c) our method.

Finally, we test the average processing time of the three methods with more than 500 images of the dataset. All methods runs in a PC whose main config are: CPU: AMD2.0, memory Size: 1G, and VC++6.0 develop environment is used. We coarsely compare the average processing time of the three methods in table 1. It is clearly shown that our method is more efficient.

Table 1. Comparison of Average processing time

Methods	Average time(second)
Ref.[9]	40.73
Ref[10]	20.21
Our methods	7.68

4 Conclusion

VP detection is very important and useful for road detection, most VP detection methods are proposed for visual light (VL) images in last decades and many have been demonstrated very effective for road detection. However, little literature investigates infrared image based VP detection. In essential, the VP detection methods based on VL image usually can not be used to infrared images straightforward since the texture of infrared image is very weak. This paper proposed a simple but effective approach for infrared based VP detection. Experiments on real infrared road image sequences demonstrate its effectiveness and efficiency. In the future, we will apply our VP detection method to detect road boundaries in infrared road images.

Acknowledgements. Firstly, the authors would like to thank all the reviewers for their helpful comments. This work was jointly supported by National Science Funds of China under grants 61175082 and 61203246, the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety (Nanjing University of Science and Technology), Grant No. 30920130122006, Jiangsu science & technology pillar program under grants BE2011192, and technology foundation for selected overseas Chinese scholar and ministry of personnel of china.

References

1. Lookingbill, A., Rogers, J., Lieb, D., Curry, J., Thrun, S.: reverse optical flow for self-supervised adaptive autonomous robot navigation. *International Journal of Computer Vision* 74(3), 287–302 (2007)
2. Thorpe, C., Hebert, M., Kanade, T., Shafer, S.: vision and navigation for the Carnegie-Mellon Navlab. *IEEE Trans. Pattern Analysis and Machine Intelligence* 10(3), 362–373 (1988)
3. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection: A review. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 28(5), 694–711 (2006)
4. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 34(4), 743–761 (2012)
5. Kong, H., Autibert, J.-Y., Ponce, J.: General Road detection from a single image. *IEEE Transaction on Image Processing* 19(8), 2211–2220 (2010)
6. Ren, M., Sun, H., Wang, H., Yang, J.: Road Boundary detection in infrared image for ALV navigation. *Infrared and Laser Engineering* 35(1), 106–110 (2006)
7. Alvarez, J.M., Gevers, T., Lopez, A.M.: *IEEE International Conference on Computer Vision and Pattern Recognition*. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2010)
8. Kong, H., Autibert, J.-Y., Ponce, J.: Vanishing point detection for road detection. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2009)
9. Moghadam, P., Starzyk, J.A., Wijesoma, W.S.: Fast vanishing-point detection in unstructured environment. *IEEE Transaction on Image Processing* 21(1), 425–430 (2012)
10. Kong, H., Sarma, S.E., Tang, F.: Generalizing laplacian of Gaussian Filters for vanishing point detection. *IEEE Transactions on Intelligent Transportation Systems* (2013)
11. Siagian, C., Chang, C.-K., Itti, L.: Mobile robot navigation system in outdoor pedestrian environment using vision-based road recognition. In: *IEEE International Conference on Robotics and Automation, ICRA* (2013)
12. Buades, A., Coll, B., Morel, J.-M.: A non-local algorithm for image denoising. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2005)
13. Wang, S., Zhang, L., Liang, Y.: Nonlocal spectral prior model for low-level vision. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) *ACCV 2012, Part III. LNCS*, vol. 7726, pp. 231–244. Springer, Heidelberg (2013)
14. Huang, K., Wang, L., Tan, T., Maybank, S.: A real-time object detecting and tracking system for outdoor night surveillance. *Pattern Recognition* 41(1), 432–444 (2008)

Neuro-control to Energy Minimization for a Class of Chaotic Systems Based on ADP Algorithm

Ruihuo Song^{1,*}, Wendong Xiao¹, and Qinglai Wei²

¹ School of Automation and Electrical Engineering,
University of Science and Technology Beijing, Beijing, 100083, China
ruizhuosong@163.com

² The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

Abstract. This paper discusses the energy minimization problem of a class of chaotic systems, and constructs an optimal neuro-controller based on adaptive dynamic programming (ADP) algorithm. To learn the optimal performance index and control policy, an iterative algorithm is established. To prove the convergence of the presented iterative algorithm, theorems with rigorous and detailed proofs are given. It is proven that the iterative performance index functions are monotone decreasing and converge to the minimum energy. A simulation example is used to indicate that the presented energy minimization control method is effective.

Keywords: Energy minimization, adaptive dynamic programming, chaotic systems, approximate dynamic programming, optimal control.

1 Introduction

During the past decades, various problems on chaotic systems are discussed [1,2]. In the case of controller designing for the chaotic systems, many methods can be seen in [4,5], from theory to practice. The popular methods include impulsive control method [6,7], adaptive synchronization control method [3], etc. However, people pursue to spending minimum energy and obtaining satisfying results, in recent years. So in [8], for a class of discrete-time chaotic systems, an optimal tracking control is designed based on approximation-error-ADP algorithm, which opens up the new directions for chaotic optimal control.

For optimal control problems, we all know that the dynamic programming is an useful tool [9,10,11,12]. In the early 1970's, Werbos established adaptive dynamic programming (ADP), which can obtain faster and better performance than dynamic programming [13]. In the past few years, Jagannathan [14], Powell [15], [11], Lewis [17], [18], [19], Murray [9], Liu [21,22], Si [23,24] have developed ADP algorithms. [25] defined a novel performance index function for the optimal

* Corresponding author.

tracking problem. In addition, [26] proposed a new optimal control algorithm for systems with control constraints. Aiming at the unknown system dynamics, [27] proposed a forward-in-time method to obtain the optimal control solution. Furthermore, [28] presented an optimal control method for unknown nonlinear systems. Based on the abundant research results, we will present the optimal control for chaotic systems, which is used to realize the energy minimization. First, the studied problem of continuous time chaotic systems is given. Then the minimum energy performance index function is proposed, and the corresponding optimal control is also presented. For obtaining the optimal solution, an iterative algorithm based on ADP is estimated. Next, as the convergence of the presented iterative algorithm is an important issue which need to be analysed. So some theorems are given to support the iterative algorithm. At last, a simulation is necessary to verify the optimal method for chaotic systems.

The rest of this paper is organized as follows. In Section 2, we present the problem formulation. In Section 3, the optimal control method is developed and the convergence proof is given. In Section 4, an example is given to demonstrate the effectiveness of the proposed control scheme. In Section 5, the conclusion is given.

2 Problem Formulation

Consider a class of controlled chaotic systems

$$\dot{x} = f(x) + Hu(x), \quad (1)$$

where $x(t)$ is n -dimension state vector, $u(t)$ is m -dimension control vector, $f(x(t)) \in \mathbb{R}^n$ is considered as smooth functions, $H \in \mathbb{R}^{n \times m}$ is constant matrix. Assume that the chaotic system is controllable.

Define the following quadratic performance index

$$V(x_0) = \int_0^\infty [x^T Q x + u^T R u] dt, \quad (2)$$

where Q, R are symmetric positive definite.

The energy minimization control problem is to find the optimal control policy $\{u(x)\}$, such that the performance index (2) is minimal subject to system (1). Such controls are defined to be admissible. For convenience, $V^*(x)$ is used to denote the optimal performance index which is defined as $V^*(x) = \min_u V(x)$, and u^* is used to denote the corresponding optimal control policy.

Definition 1. A control $u(x)$ is defined to be admissible with respect to (1) on \mathbb{R}^m if $u(x)$ is continuous on \mathbb{R}^m , $u(0) = 0$, $u(x)$ stabilizes (1) and $\forall x(0) \in \mathbb{R}^n$, $V(x(0))$ is finite.

3 Optimal Control Method

As everyone knows, Eq. (2) can be expanded as follows:

$$V(x_0) = \int_0^T [x^T Q x + u^T R u] dt + V(x(T)). \quad (3)$$

So we can obtain

$$\lim_{T \rightarrow 0} \frac{1}{T} (V(x_0) - V(x(T))) = \lim_{T \rightarrow 0} \frac{1}{T} \int_0^T [x^T Q x + u^T R u] dt. \quad (4)$$

That is

$$\dot{V} = -x^T Q x - u^T R u. \quad (5)$$

So we get the well-known HJB equation as follows

$$\begin{aligned} & (V_x^*)^T (f(x) + H u(x)) + x^T Q x + u^T (x) R u(x) \\ &= (V_x^*)^T f(x) + x^T Q x - \frac{1}{4} (V_x^*)^T H R^{-1} H^T V_x^* = 0 \\ & V_x^*(0) = 0, \end{aligned} \quad (6)$$

and the corresponding optimal control policy

$$u^*(x) = -\frac{1}{2} R^{-1} H^T V_x^*. \quad (7)$$

It is clear that the energy minimization control problem for chaotic systems can be solved if the optimal performance index $V^*(x)$ can be obtained from (6). However, there is currently no quite effective method for solving the optimal performance index. Therefore, in the following part we will discuss how to utilize the ADP iteration algorithm to obtain the solution for energy minimization control problem.

Firstly, the following Lemma shows how Eq. (7) can be used to improve the control policy.

Theorem 1. If $V^{[i]}$ satisfies the equation

$$\begin{aligned} & \left(V_x^{[i]} \right)^T \left(f(x) + H u^{[i]}(x) \right) + x^T Q x + \left(u^{[i]} \right)^T (x) R u^{[i]}(x) = 0 \\ & V^{[i]}(0) = 0, \end{aligned} \quad (8)$$

then the new control derived as

$$u^{[i+1]}(x) = -\frac{1}{2} R^{-1} H^T (x) V_x^{[i]} \quad (9)$$

is an admissible control for (1).

Proof: Since $V^{[i]}$ is positive definite it attains a minimum at the origin, and thus, $\frac{dV_x^{[i]}}{dx}$ must vanish there. This implies that $u^{[i+1]}(0) = 0$.

Moreover taking the derivative of $V^{[i]}$ along the system $f(x) + Hu^{[i+1]}(x)$, we have

$$\dot{V}^{[i]}(x, u^{[i+1]}) = \left(V_x^{[i]}\right)^T \left(f(x) + Hu^{[i+1]}(x)\right). \quad (10)$$

From (8), we know that

$$\left(V_x^{[i]}\right)^T f(x) = -\left(V_x^{[i]}\right)^T Hu^{[i]}(x) - x^T Qx - \left(u^{[i]}\right)^T(x) Ru^{[i]}(x). \quad (11)$$

Take (11) into (10), we can get

$$\begin{aligned} \dot{V}^{[i]}(x, u^{[i+1]}) &= -\left(V_x^{[i]}\right)^T Hu^{[i]}(x) - x^T Qx - \left(u^{[i]}\right)^T(x) Ru^{[i]}(x) \\ &\quad + \left(V_x^{[i]}\right)^T Hu^{[i+1]}(x). \end{aligned} \quad (12)$$

From (9), we can have

$$\left(V_x^{[i]}\right)^T H = -2 \left(u^{[i+1]}(x)\right)^T R. \quad (13)$$

So we get

$$\begin{aligned} \dot{V}^{[i]}(x, u^{[i+1]}) &= -2 \left(u^{[i+1]}(x)\right)^T R \left(u^{[i+1]}(x) - u^{[i]}(x)\right) \\ &\quad - x^T Qx - \left(u^{[i]}(x)\right)^T Ru^{[i]}(x). \end{aligned} \quad (14)$$

Based on matrix decomposition method, $R = P\Lambda P$ where Λ is a diagonal matrix with its values being the singular values of R and P is an orthogonal symmetric matrix. Substituting for R in (14) we get

$$\begin{aligned} \dot{V}^{[i]}(x, u^{[i+1]}) &= -2 \left(u^{[i+1]}(x)\right)^T P\Lambda P \left(u^{[i+1]}(x) - u^{[i]}(x)\right) \\ &\quad - x^T Qx - \left(u^{[i]}(x)\right)^T P\Lambda P u^{[i]}(x). \end{aligned} \quad (15)$$

Using the coordinate change $u = p^{-1}y$, we have

$$\begin{aligned} \dot{V}^{[i]}(x, u^{[i+1]}) &= -2 \left(p^{-1}y^{[i+1]}\right)^T P\Lambda P \left(p^{-1}y^{[i+1]} - p^{-1}y^{[i]}\right) \\ &\quad - x^T Qx - \left(p^{-1}y^{[i]}\right)^T P\Lambda P p^{-1}y^{[i]} \\ &= -x^T Qx - \sum_{k=1}^m A_{kk} \left(\left(y^{[i+1]}\right)^T y^{[i+1]}\right. \\ &\quad \left. + \left(y^{[i+1]} + y^{[i]}\right)^T \left(y^{[i+1]} + y^{[i]}\right)\right) \\ &< 0. \end{aligned} \quad (16)$$

Since $V^{[i]}(x)$ is regarded as Lyapunov function for $u^{[i+1]}$, from Definition 1, $u^{[i+1]}$ is admissible.

The proof is completed.

Theorem 2. If $V^{[i+1]}$ satisfies equation

$$\begin{aligned} \left(V_x^{[i+1]}\right)^T \left(f(x) + Hu^{[i+1]}\right) + x^T Qx + \left(u^{[i+1]}\right)^T Ru^{[i+1]} &= 0, \\ V^{[i+1]}(0) &= 0, \end{aligned} \quad (17)$$

then

$$V^*(x) \leq V^{[i+1]}(x) \leq V^{[i]}(x), \forall x. \quad (18)$$

Proof:

From (17), we can get

$$\left(V_x^{[i+1]}\right)^T f(x) = -\left(V_x^{[i+1]}\right)^T Hu^{[i+1]} - x^T Qx - \left(u^{[i+1]}\right)^T Ru^{[i+1]}. \quad (19)$$

So if we take the derivative along the system $f(x) + Hu^{[i+1]}$, according to (11) and (13), the following relation can be obtained

$$\begin{aligned} &V^{[i+1]}(x_0) - V^{[i]}(x_0) \\ &= 2 \left(u^{[i+1]}\right)^T Ru^{[i]} - \left(u^{[i]}\right)^T Ru^{[i]} \\ &\quad - \left(u^{[i+1]}\right)^T Ru^{[i+1]}. \end{aligned} \quad (20)$$

In the same method, $R = PAP$ and $u = P^{-1}y$, (20) can be improved as follow

$$\begin{aligned} &V^{[i+1]}(x_0) - V^{[i]}(x_0) \\ &= - \sum_{k=1}^m \Lambda_{kk} \left(\left(y^{[i+1]}\right)^T y^{[i+1]} + \left(y^{[i]}\right)^T y^{[i]} \right. \\ &\quad \left. - 2 \left(y^{[i+1]}\right)^T y^{[i]} \right) \\ &= - \sum_{k=1}^m \Lambda_{kk} \left(y^{[i+1]} - y^{[i]} \right)^T \left(y^{[i+1]} - y^{[i]} \right) \\ &\leq 0. \end{aligned} \quad (21)$$

Moreover, from the definition of V^* , we know that $V^*(x) \leq V^{[i+1]}(x), \forall x$. Thus, $V^*(x) \leq V^{[i+1]}(x) \leq V^{[i]}(x), \forall x$.

The proof is completed.

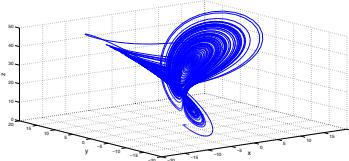


Fig. 1. The chaotic system trajectory

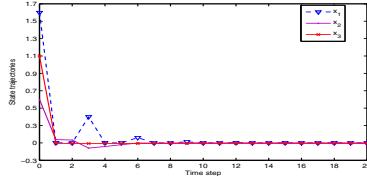


Fig. 2. The state trajectories x , y and z

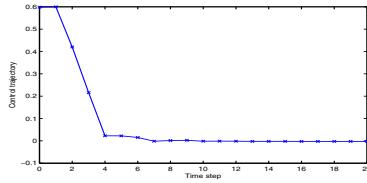


Fig. 3. The control trajectory

4 Simulation Study

In this section, we evaluate the performance of our energy minimization control scheme by applying the method on Lorenz chaotic system. The Lorenz system is described as follows

$$\begin{cases} \dot{x} = -a(x - y), \\ \dot{y} = rx - y - xz + u, \\ \dot{z} = -bz + xy, \end{cases} \quad (22)$$

where x , y and z are state variables and a , b and r are non-negative known constants. In the case that $a = 10$, $b = \frac{8}{3}$ and $r = 28$ system exhibits a chaotic behavior as shown in Fig. 1. Using the presented optimal control method, we get the simulation results. In Fig. 2, the state trajectories are shown, which is asymptotically stable as time growth. The control trajectory is shown in Fig. 3. It is also convergent. From the simulation, we can see that the optimal algorithm is very effective and feasible.

5 Conclusions

This paper presents an optimal method to minimize energy requirements. For a class of chaotic systems, an energy minimization performance index function is

given. To get the optimal one, an iterative ADP algorithm is used. It can get the optimal solution approximatively. For verifying the effectiveness of the presented iterative algorithm, corresponding theorems are given, rigorously. The simulation is used to demonstrate the feasibility of the presented optimal method.

Acknowledgment. This work was supported in part by the Open Research Project from SKLMCCS (Grant No. 20120106), the Fundamental Research Funds for the Central Universities (Grant No. FRF-TP-13-018A), the China Postdoctoral Science Foundation (Grant No. 2013M530527), the National Natural Science Foundation of China (Grants No. 61304079, 61374105), and Beijing Natural Science Foundation (Grant No. 4132078).

References

1. Zhang, H., Huang, W., Wang, Z., Chai, T.: Adaptive synchronization between two different chaotic systems with unknown parameters. *Physics Letters A* 350(5-6), 363–366 (2006)
2. Chen, S., Lü, J.: Synchronization of an uncertain unified chaotic system via adaptive control 14(4), 643–647 (2002)
3. Zhang, H., Wang, Z., Liu, D.: Chaotifying fuzzy hyperbolic model using adaptive inverse optimal control approach 14(10), 3505–3517 (2004)
4. Ma, T., Fu, J.: Global exponential synchronization between L system and Chen system with unknown parameters and channel time-delay. *Chinese Physics B* 20(5), 050511 (2011)
5. Ma, T., Fu, J., Sun, Y.: An improved impulsive control approach to robust lag synchronization between two different chaotic systems. *Chinese Physics B* 19(9), 090502 (2010)
6. Ma, T., Zhang, H., Fu, J.: Exponential synchronization of stochastic impulsive perturbed chaotic Lur'e systems with time-varying delay and parametric uncertainty. *Chinese Physics B* 17(12), 4407–4417 (2008)
7. Zhang, H., Ma, T., Fu, J., Tong, S.: Robust lag synchronization of two different chaotic systems via dual-stage impulsive control. *Chinese Physics B* 18(9), 3751–3757 (2009)
8. Song, R., Xiao, W., Sun, C., Wei, Q.: Approximation-Error-ADP-Based Optimal Tracking Control for Chaotic Systems With Convergence Proof. *Chinese Physics B* (accept)
9. Murray, J., Cox, C., Lendaris, G., Saeks, R.: Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 32(2), 140–153 (2002)
10. Seiffertt, J., Sanyal, S., Wunsch, D.: Hamilton-Jacobi-Bellman equations and approximate dynamic programming on time scales. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(4), 918–923 (2008)
11. He, P., Jagannathan, S.: Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 37(2), 425–436 (2007)
12. Zhao, Y., Patek, S., Beling, P.: Decentralized Bayesian search using approximate dynamic programming methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(4), 970–975 (2008)

13. Werbos, P.: Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook* 22, 25–38 (1977)
14. Zheng, C., Jagannathan, S.: Generalized Hamilton-Jacobi-Bellman formulation-based neural network control of affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks* 19(1), 90–106 (2008)
15. Powell, W.: Approximate dynamic programming: solving the curses of dimensionality. Wiley, New York (2009)
16. He, P., Jagannathan, S.: Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 37(2), 425–436 (2007)
17. Al-Tamimi, A., Lewis, F., Abu-Khalaf, M.: Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica* 43, 473–481 (2007)
18. Vrabie, D., Pastravanu, O., Abu-Khalaf, M., Lewis, F.: Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica* 45(2), 477–484 (2009)
19. Vamvoudakis, K., Lewis, F.: Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica* 46(5), 878–888 (2010)
20. Murray, J., Cox, C., Lendaris, G., Saeks, R.: Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 32(2), 140–153 (2002)
21. Wang, F., Jin, N., Liu, D., Wei, Q.: Adaptive dynamic programming for finite horizon optimal control of discrete-time nonlinear systems with ε -error bound. *IEEE Transactions on Neural Networks* 22(1), 24–36 (2011)
22. Zhang, H., Wei, Q., Liu, D.: An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games. *Automatica* 47(1), 207–214 (2011)
23. Si, J., Wang, Y.: On-line learning control by association and reinforcement. *IEEE Transactions on Neural Networks* 12(2), 264–276 (2001)
24. Enns, R., Si, J.: Helicopter trimming and tracking control using direct neural dynamic programming. *IEEE Transactions on Neural Networks* 14(4), 929–939 (2003)
25. Zhang, H., Wei, Q., Luo, Y.: A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(4), 937–942 (2008)
26. Zhang, H., Luo, Y., Liu, D.: Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks* 20, 1490–1503 (2009)
27. Wei, Q., Zhang, H., Dai, J.: Model-free multiobjective approximate dynamic programming for discrete-time nonlinear systems with general performance index functions. *Neurocomputing* 72(7-9), 1839–1848 (2009)
28. Wang, D., Liu, D., Wei, Q., Zhao, D.: Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica* 48(8), 1825–1832 (2012)

Low SNR FMCW Signal Processing with Prior Information

Qing Wen Hou, Zhi Wei Xu, Zhen Long Bai, Xian Zhong Chen, and Jing Ni Wang

School of Automation and Electrical Engineering, University of Science and Technology Beijing
houqw@ustb.edu.cn

Abstract. Aiming at the range ambiguity problem of wideband and low SNR (signal-to-noise ratio) FMCW (frequency modulated continuous wave) signals under severe measurement conditions, a solid surface FMCW signal model is built. Besides, based on prior information, a distance inversion method is proposed, which contains signal pretreatment, band intercepting and energy weighting. Actual data processing results show that, SNR of the frequency spectrum has been raised from -5.55dB to 8.32dB after pretreatment; and the final results show that index of similarity of stock level shape respectively raised from 0.46 to 0.71 and 0.79.

Keywords: FMCW radar, prior information, low SNR, processing algorithm, distance measurement.

1 Introduction

In a closed container, material height and shape are the important parameters in the industrial production. Microwave is sensitive to material surface asperity, and is not affected by temperature, dust, gas and pressure. So microwave (e.g. FMCW) radar has been widely used in industrial production, especially in some severe conditions[1].

Radar imaging could be divided into two categories according to the antenna arrangement and signal processing. The first method is suitable for regular array radars' data. It interpolates after 2D-FFT, and then gets 3D image from 2D synthetic aperture. Although its image is accurate, its theory is complex and it takes a long time. The second method is suitable for irregular single radar's data. The material shape could be figured from calculated height data approximately. Its theory is simple, but its image is not accurate enough due to sparse data.

MIMO-based radar imaging system[2] belongs to the first method. It is flexible, but complex. Furthermore, single radar is smaller than array radar, so it is more suitable for some industrial field (e.g. blast furnace). However, the height data with noise used in second method are not credible enough. In order to find the solution, scholars put forward many methods. Liu Xin'en[3] combined response surface model of Gaussian process and Monte Carlo method. The credible data could be selected by a standard Latin hypercube according to the known probability distribution of input variables. Aiming at the problem of surface reconstruction from sparse data, Xu Shouqian[4] classified the points according

to their features, and introduced iteration and asymptotics to make up data. Li Leqing[5] used the RBF (Radial Basis Functions) neural network and the correlation coefficient to self-adjust the radius of the kernel function. Yutaka Otake[6] combined an adaptive partition of unity approximation with least-squares RBF fitting to generate a high quality surface reconstruction. Keith Biggers[7] used a set of user-provided models for prior knowledge and applied this knowledge to the iterative identification and construction process among a cluttered scene.

This paper uses stock distributing experience and actual distributing mode for prior information and applies this information to correct surface fitting.

2 Solid Surface FMCW Signal Model

Based on multiple data analysis to invert surface, three models can be built and signals will be classified by angle of incidence θ , signal-to-noise ratio snr and maximum signal energy a . In the following expressions, \overline{SNR} is the mean SNR, and \overline{A} is the mean maximum energy of signals.

1. Model A:

$$\begin{cases} a \leq \overline{A} \text{ or } snr \leq \overline{SNR} \\ \theta > 30^\circ \end{cases} \quad (1)$$

2. Model B:

$$\begin{cases} a > \overline{A} \\ snr > \overline{SNR} \text{ or } \theta > 30^\circ \end{cases} \quad \begin{cases} a < \overline{A} \\ snr < \overline{SNR} \\ \theta \leq 30^\circ \end{cases} \quad (2)$$

3. Model C:

$$\begin{cases} a \geq \overline{A} \text{ or } snr \geq \overline{SNR} \\ \theta \leq 30^\circ \end{cases} \quad (3)$$

In model A, the incident angle θ is greater than 30° . Because of the influence of backscattering, echo intensity in model A is far less than model C which has a small incident angle. What's more, due to the complex and obvious noise in the actual measurement environment, the spectrum of large incident angle shows low SNR, wide band and fuzzy distance information.

Different SNR makes credibility different. The credibility is lacking in the model which has low SNR and large incident angle. So model C has the highest credibility, its results would be used directly, while model A has the lowest credibility, its results should be replaced with symmetric points. The credibility of model B should be analyzed in combination with other conditions (e.g. neighbor or symmetric points).

3 Large Angle and Low SNR Signal Processing with Prior Information

In order to improve low SNR and range ambiguity caused by the spectrum of continuous multiple peaks, this paper presents a signal processing inversion method to

compute the central distance. This method denoises by using prior information and deblurs by using the energy center of gravity thought. In the processing of real signals, the priori information can be used to identify, count and optimize data.

3.1 Signal Pretreatment

In the actual measurement, the background noise is all over the entire frequency axis including the effective signal frequency range, and cannot be removed by digital filter. The background noise will be calculated respectively in different frequency ranges by identifying target echo with prior information. Its value in a certain frequency range can be estimated combining with the effective surface echo data in the same range. It supposed that N sets of data are measured, do 1024 points FFT on them. N sets of spectrum data synthesize a $N \times 512$ matrix named S .

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix} = \begin{bmatrix} S_1(1) & S_1(2) & \cdots & S_1(512) \\ S_2(1) & S_2(2) & \cdots & S_2(512) \\ \vdots & \vdots & & \vdots \\ S_N(1) & S_N(2) & \cdots & S_N(512) \end{bmatrix}. \quad (4)$$

FMCW beat signal frequency is proportional to distance, so spectrum data above are corresponding to the certain distances. The range of distance is from R_{\min} to R_{\max} . Let $R_1 = \frac{R_{\min} + R_{\max}}{3}$, $R_2 = \frac{2(R_{\min} + R_{\max})}{3}$, the distances $R_{\min}, R_1, R_2, R_{\max}$ correspond with the frequencies (i.e. spectral lines) L_1, L_2, L_3, L_4 respectively.

Find the data sets $N1, N2, N3$ which the numbers of elements are $n1, n2$ and $n3$. Make each set of data only contain background noise in corresponding distance range $R_{\min} \sim R_1, R_1 \sim R_2, R_2 \sim R_{\max}$. The noise spectrum can be obtained as follows:

$$Noise(j) = \begin{cases} \frac{\sum_{i=1}^N S_i(j)}{N} & j \in [1, L_1] \cup [L_4, 512] \\ \frac{\sum_{i=1}^{n1} S_{N1(i)}(j)}{n1} & j \in (L_1, L_2) \\ \frac{\sum_{i=1}^{n2} S_{N2(i)}(j)}{n2} & j \in [L_2, L_3] \\ \frac{\sum_{i=1}^{n3} S_{N3(i)}(j)}{n3} & j \in [L_3, L_4] \end{cases}. \quad (5)$$

After preprocessing, signal spectrum of group i can be expressed as:

$$s_i(j) = \begin{cases} S_i(j) - Noise(j), & S_i(j) > Noise(j) \\ 0, & S_i(j) \leq Noise(j) \end{cases}. \quad (6)$$

j is spectral line number, $1 \leq j \leq 512$.

3.2 Spectrum Zooming and Interception

For traditional distance inversion method, its frequency resolution of FFT spectrum is as low as f_s/N (f_s is sampling frequency). CZT (Chirp-Z Transform) is suitable for detailed analysis on the narrow band spectrum, and can improve the accuracy effectively [8].

The signal spectrum has multiple peaks and wide band. By intercepting spectrum, band with effective distance information can be received. Make full use of the whole information, result will be more accurate. Drawing envelope or least square fitting curve on CZT spectrum helps band interception.

Envelope method can obtain dominant and prominent characteristics of curve change by connecting the peaks. And least squares fitting method can get smooth curve of the minimum error compared with the original CZT spectrum. It combines all the information of points, reflects the trend of curve.

Find two points satisfied the following conditions as the left and the right ends of band, their coordinates are $(f(i), a(i))$ and $(f(j), a(j))$. Let the coordinate of the spectrum peak be $(f(n0), a(n0))$, then:

$$\begin{cases} f(i) < f(n0) \\ f(i) \leq \forall f(k) \quad i-3 \leq k \leq i+3 \\ a(i) \leq 0.2 \cdot a(n0) \end{cases} \quad (7)$$

Select the point whose frequency is closest to $f(n0)$ in the points satisfied the conditions above as the left end of band. Similarly as the right end of band.

3.3 Energy Weighting for Distance

Aiming at the problem of fuzzy central distance caused by the broadband spectrum, the method of energy weighting can be adopted to deblur in the truncated band. Central frequency f_0 can be calculated as follows:

$$f_0 = \frac{\sum_{i=n_{\min}}^{n_{\max}} A_i^2 \cdot f_i}{\sum_{i=n_{\min}}^{n_{\max}} A_i^2} \quad (8)$$

n_{\min}, n_{\max} : The corresponding spectral line number to the frequency of the closest point to the left/right end of band in the CZT spectrum;

A_i : Amplitude of points in the CZT spectrum;

f_i : The corresponding frequency of each point in the truncated CZT spectrum.

According to the working principle of FMCW:

$$f = \frac{2R}{c} \cdot \Delta f \cdot f_s \quad (9)$$

Δf : FMCW stepped frequency;

c : The propagation velocity of the electromagnetic wave.

So the central distance R_0 is:

$$R_0 = f_0 \cdot \frac{c}{2 \cdot f_s \cdot \Delta f} = \frac{c \cdot \sum_{i=n_{\min}}^{n_{\max}} A_i^2 \cdot f_i}{2f_s \Delta f \sum_{i=n_{\min}}^{n_{\max}} A_i^2}. \quad (10)$$

The method described in section 3 has higher resolution than FFT method. Because the energy is weighted in a certain frequency range, it effectively increases and emphasizes the weighting of the large spectrum peak in the calculation of central frequency. Compared with the envelope method, the least square method intercepts wider bandwidth, and the resulting curve is smoother.

4 The Results of the Actual Signal Processing

Measuring environment is shown as figure 1. The angle θ starts from 65° , the incident angle reduces one degree each measurement time, and 60 groups of data will be measured. According to the prior information, it can be known that the surface shape is basically symmetric about the center. In model A, the small backscattering coefficient caused by the large incident angle will decrease signal intensity, even bring the furnace wall reflection effect. It is not conducive to spectrum analysis, so that the advantage of basically symmetric surface is available to analyze the point's height.

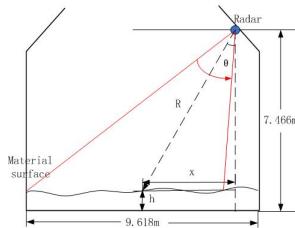


Fig. 1. Test condition of FMCW signal data

For the data of group N ($1 \leq N \leq 60$):

$$\begin{cases} \theta = 66 - N \\ x = R \cdot \sin\left(\frac{\theta \cdot \pi}{180}\right) \\ h = 7.466 - R \cdot \cos\left(\frac{\theta \cdot \pi}{180}\right) \end{cases}. \quad (11)$$

4.1 Pretreatment

Except for the fixed interference of 2m and 25m, the spectrum of model A displays the existence of two echo spectrums in 11-13m. The reason is that the large incident angle makes microwave enter near the junction of the edge of surface and the furnace wall, so there is the furnace wall reflection. Most of the incident waves reflect directly after backscattering, small part of them reflect through the surface to the furnace wall and are received by the receiving module. During this time, the distance increases, so the spectrum shows two peaks (the left peak represents the actual distance).

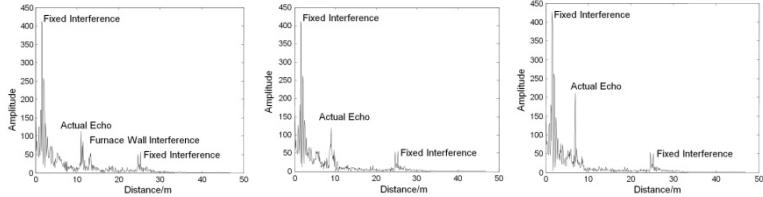


Fig. 2. Original spectrums of model A, B and C

After signal preprocessing and removing the background noise, the signal spectrums are shown as below:

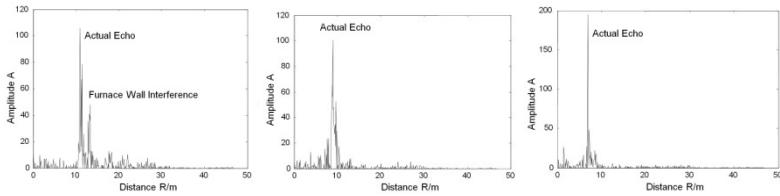


Fig. 3. Spectrums of model A, B and C after signal pretreatment

Table 1. Incident angle and SNR of signals

	Model A	Model B	Model C
Angle of incidence /°	62	39	9
SNR (Original) /dB	-7.4035	-4.9204	-3.4556
SNR (After pretreatment) /dB	3.6371	12.3996	7.2781

When the incident angle is from 56° to 60° , the actual echo cannot be distinguished, because of the strong reflection from the furnace wall. By removing these five groups of data with the lowest SNR, the average SNR is raised from -5.5544 dB to 8.3238 dB after pretreatment.

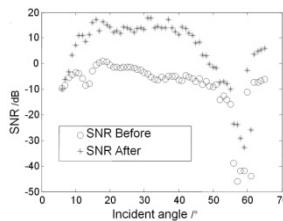
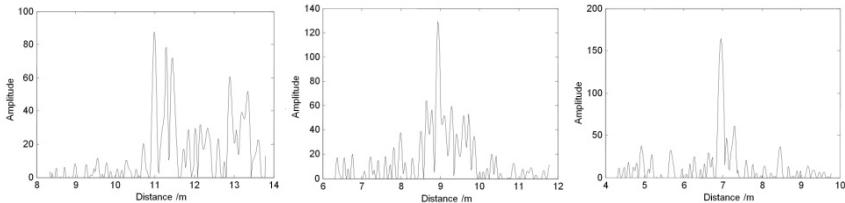


Fig. 4. SNR (before and after pretreatment)

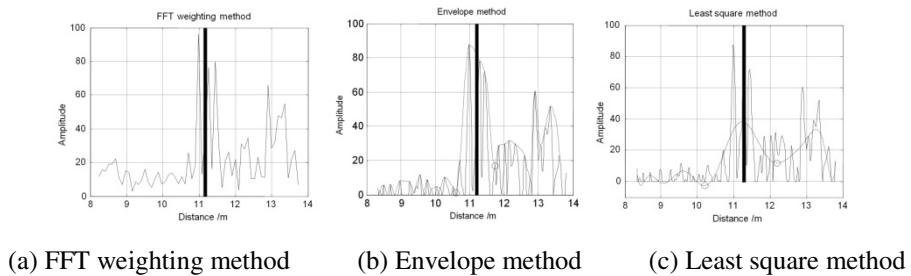
4.2 Spectrum Zooming

The CZT spectrums of the signal without noise are shown as figure 5. The range resolution is improved from 0.0916 to 0.0183 after spectrum zooming.

**Fig. 5.** CZT spectra of model A, B and C

4.3 Band Interception and Calculation of Central Distance

Take the data of model A for instance, figure 6 shows the comparison of three methods in frequency domain (a bold black line represents the central distance): (a) FFT weighting method (the mean value of the distances corresponded to the current three peaks); (b) Envelope method; (c) Least square method.



(a) FFT weighting method (b) Envelope method (c) Least square method

Fig. 6. Distance contrast of different methods**Table 2.** Central distance contrast of different methods

	Model B	Model C	Model A
FFT weighting inversion method /m	11.1904	9.1317	7.0704
Envelope inversion method /m	11.2105	9.0579	6.9882
Least square inversion method /m	11.2790	9.0519	6.9792
Actual distance /m	11.45	8.67	7.01

4.4 Stockline Fitting

The distance results of 60 groups of data are interpolated after coordinate change. After coordinate change, the distances are about 10 meters between the radar and the points whose angle of incidence are 61° to 65° . But the diameter of the blast furnace is 9.618 meters. It is visible that the results are not accurate, because of the furnace wall interference. The surface is almost symmetric due to the fabric rules, so the height of unreliable data of the large incident angle can be approximately represented by the symmetric data of the small incident angle.

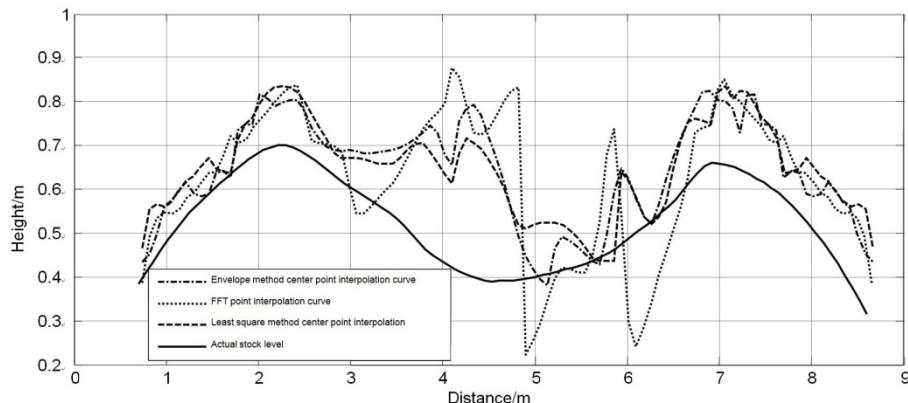


Fig. 7. Stock level contrast

Table 3. Stock level contrast

	Mean error	Shape similarity
FFT weighting inversion method	0.0964	0.4643
Envelope inversion method	0.1160	0.7110
Least square inversion method	0.1205	0.7919

Figure 7 shows that the processing results of measurement data have uniformly linear deviation of about 0.1 meters in terms of height, and it can be improved through compensation. However, two methods proposed in this paper increase by 53.13% and 70.56% respectively than the FFT weighting inversion method in terms of similarity of the surface.

5 Conclusion

In view of the complex situation of solid surface measurement, this paper proposes a low SNR FMCW signal processing method based on the priori information. This method gets the final processing results through a series of process, including denoising by the signal preprocessing, spectrum zooming, spectrum band intercepting and energy weighting. From the processing results of the actual data, it can be seen that the method proposed in this paper well smoothes the singular points after FFT, retains the original information, and gets smoother interpolation curve which is closer to the actual stock level at the same time.

References

1. Liu, J., Chen, X.Z., Zhang, Z.: A Novel Algorithm in the FMCW Microwave Liquid Level Measuring System. *Measurement Science and Technology* 7, 135–138 (2006)
2. Liu, X.M.: Application of Radar Material Level Meter in Metallurgical Production Process. *Instrument Standardization and Metrology* 1, 28–30 (2011)

3. Ko, H.H., Cheng, K.W., Su, H.J.: Range Resolution Improvement for FMCW Radars. In: the 5th European Radar Conference, pp. 352–355 (2008)
4. Hou, Q.W., Chen, X.Z., Wang, X.P.: Improved Phase-difference Algorithm with Weighted Compensation and Correction for FMCW Signal. Chinese Journal of Scientific Instrument 31, 721–726 (2010)
5. Chen, X.Z., Liu, J.: Fast Frequency Estimation Algorithm for FMCW Microwave Liquid Level Measurement. Chinese Journal of Sensors and Actuators 18, 901–905 (2005)
6. Musch, T.: A High Precision 24-GHz FMCW Radar Based on a Fractional-N Ramp-PLL. In: IEEE Transactions on Instrumentation and Measurement, pp. 324–327. IEEE Press, Alaska (2003)
7. Max, S., Vossiek, M., Gulden, P.: Fusion of FMCW Secondary Radar Signal Beat Frequency and Phase Estimations for High Precision Distance Measurement. In: The 5th European Radar Conference, pp. 124–127. IEEE Press, Amsterdam (2008)
8. Li, T.Y., Ge, L.D.: Research of Two Kinds of Fast Zoom Spectrum. System Engineering and Electronics 6, 731–733 (2004)

Sparsity Preserving Score for Joint Feature Selection

Hui Yan

Nanjing University of Science and Technology,
210094 Nanjing, China
yanhui@njust.edu.cn

Abstract. Based on recent advances in sparse representation technique, we propose in this paper Sparsity Preserving Score (SPS) to jointly select features. SPS evaluates the importance of a feature by its power of sparse reconstructive relationship preserving, which is achieved by minimizing an objective function with l_1 -norm regularization and binary constrain. Our searching strategy, which is an essentially discrete optimization, jointly selects features by projecting the original high-dimensional data to a low-dimensional space through a special binary projection matrix. Theoretical analysis guarantees our objective function can get a closed form solution, which is as simple as scoring each feature by Frobenius norm of sparse linear reconstruction residual for each feature. Comparing experiments on two face datasets are carried out. The experimental results demonstrate the effectiveness and efficiency of our algorithm.

Keywords: feature selection, sparse representation, binary projection matrix.

1 Introduction

Dimensionality reduction can be achieved by either feature selection or feature extraction [1] to a low dimensional space. In contrast to feature extraction, feature selection aims at finding out the most representative or discriminative subset of the original feature spaces according to some criteria.

Feature selection mainly focuses on search strategies and measurement criteria. The search strategies for feature selection can be divided into three categories: exhaustive search, sequential search, and random search. The exhaustive search aims to find out the optimal solution from all possible subsets. However, it is NP-hard and thus it is impractical to run. Sequential search methods, such as sequential forward selection and sequential backward elimination [9], start from an empty set or the set of all candidates as the initial subset selected, and successively add features to the selected feature or eliminate features from a subset one by one. The major drawback of the traditional sequential search methods relies heavily on search routes. Although the sequential methods do not guarantee the global optimality of selected subset, they have been widely used because of their simplicity and relatively low computational cost even for large-scale data. Plus-l-minus-r (l-r) [10], a slightly more reliable sequential search method, considers deleting features that were previously selected and selecting features that were previously deleted. However, it only partially solves the limit of search routes and bring in additional parameters. The random search methods, such as the random hill climbing and

its extension Sequential Floating Search [11], take advantage of randomized steps of the search, and select features from all candidates with a chance probability per feature.

Despite the encouraging results achieved in certain cases, the main drawback of traditional searching strategies is that they simply use statistical character to rank the features, and then select features individually. Therefore, they neglect the interaction and dependency among features. Recently, sparsity regularization has been applied to joint feature selection and subspace learning [2], [3], [4]. These researches use $l_{2,1}$ -norm on the projection matrix to achieve row-sparsity, which leads to selecting relevant features and extracting discriminant features simultaneously. Meanwhile, straightforward feature selection strategy (or projection based feature selection) [5], provides us with an effective way for joint feature selection. It selects optimal features simultaneously, which means the optimal subset consisting of $(p+1)$ features is independent of optimal subset consisting of p features.

Measurement criterion is also an important research direction in feature selection. Data Variance [6] ranks the score of each feature by the variance along a dimension. The measurement criterion of Data Variance finds features that are useful for representing data, however, these features may be not useful for preserving discriminative information. Laplacian Score [15] is a recent locality graph-based unsupervised feature selection algorithm. Laplacian score reflects locality preserving power of each feature.

Recently, Wright et al. present a Sparse Representation-based Classification (SRC) [7] method. Afterwards, sparse representation-based feature extraction becomes an active direction. Qiao et al. [12] present a Sparsity Preserving Projections (SPP) method, which aims to preserve the sparse reconstructive relationship of the data. Zhang et al. [14] recently present a graph optimization for dimensionality reduction with sparsity constraints, which can be viewed as an extension of SPP. Clemmensen et al. [13] provide a sparse linear discriminant analysis with a sparseness constraint on projection vectors.

As we know, feature selection with direct connection to SRC has not emerged. In this paper, we use SRC as a measurement criterion to design a joint feature selection algorithm called Sparsity Preserving Score (SPS). The formulated objective function, which is an essentially discrete optimization, aims to seek a binary linear transformation such that in a low-dimensional space, the sparse representation coefficients are preserved. Theoretical analysis guarantees our function can get a closed form solution, which is as simple as scoring each feature by Frobenius norm of sparse linear reconstruction residual for each feature.

2 Sparse Representation Based classification (SRC)

Let $A_i = [x_{1,1}, x_{1,2}, \dots, x_{1,N_i}]$ be the set of the training samples from Class i , where N_i is the number of training samples of Class i , and $A = [A_1, A_2, \dots, A_K]$ be the set composed of entire training samples from K classes, where $N_1 + N_2 + \dots + N_K = N$. For a new test sample y , SRC computes its sparse representation $\hat{\alpha}$:

$$\hat{\alpha} = \arg \min \|\alpha\|_1, \text{ s.t. } A\alpha = y \quad (1)$$

If y is from Class i , ideally, the nonzero entries in α will all be associated with the columns of A from a single class i . In the other words, the sparse non-zero entries in α can well encode the identity of the test sample y [7]. Using only the coefficients associated with the i -th class, i.e., $\delta_i(\hat{\alpha})$, we compute the residual between y and its prototype of Class i as

$$r_i(y) = \|y - \delta_i(\hat{\alpha})\|_2$$

If $r_k(y) = \min_i r_i(y)$, SRC assigns y to class k .

3 Sparsity Preserving Score

We formulate our strategy of feature selection as follows: given a set of unlabeled training samples $x_i \in R^m, i = 1, \dots, N$, learn a transformation matrix (or feature selection matrix) $P \in R^{n \times m} (n < m)$ such that P is optimal according to our objective function. Specially, the feature selection algorithm requires P being a 0-1 binary matrix, i.e., $P(i, k_i) = 1 (i = 1, \dots, n, k_i \leq m)$ and the rest entries should be equal to zero. In other words, the set of k_i is the sequence numbers of the selected features. Obviously, the matrix P should satisfy two constraints: (1) each row of P has one and only one non-zero entry of 1; (2) each column of P has at most one non-zero entry. Accordingly, the sum of entries in each row equals 1 and the sum of entries in each column less than or equals 1. We define the following objective function to minimize the sparse linear reconstruction residual, and measure the sparsity by the l_1 -norm of coefficients.

$$\min_{P, \{\beta_i, i=1, \dots, N\}} \sum_{i=1}^N \|Px_i - PD_i\beta_i\|_F^2 + \lambda \|\beta_i\|_1 \quad (2)$$

$$\begin{aligned} \text{s.t. } & \sum_{j=1}^m P(i, j) = 1 \\ & \sum_{i=1}^n P(i, j) \leq 1 \\ & P(i, j) = 0 \quad \text{or} \quad 1 \end{aligned}$$

Here $D_i = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N] \in R^{m \times (N-1)}$ is the collection of training samples without the i -th sample, β_i is the sparse representation coefficient vector of x_i over D_i and λ is a scalar parameter. The items in line 1 of (2) are approximation and sparse constraints in the features selected space respectively. (2) is a joint optimization of P and $\beta_i, i = 1, \dots, N$.

Let

$$J_{P, \{\beta_i, i=1, \dots, N\}} = \sum_{i=1}^N \|Px_i - PD_i\beta_i\|_F^2 + \lambda \|\beta_i\|_1 \quad (3)$$

The minimization of $J_{P, \{\beta_i, i=1, \dots, N\}}$ can be implemented by optimizing P and β_i respectively.

Firstly, we compute β_i in the original space as

$$\min_{\beta_i, i=1, \dots, N} \sum_{i=1}^N \|x_i - D_i\beta_i\|_F^2 + \lambda \|\beta_i\|_1 \quad (4)$$

Some standard convex optimization techniques or TNIPM in [8] can be used to solve β_k . In our experiments, we directly use source code provided by authors in [8].

After then, let $\Gamma = [\gamma_1, \dots, \gamma_N]$, $\gamma_i = x_i - D_i\beta_i$, the objective function(3) is reduced to

$$\min_P \text{trace}\{P\Gamma\Gamma^TP^T\} \quad (5)$$

$$\begin{aligned} \text{s.t. } & \sum_{j=1}^m P(i, j) = 1 \\ & \sum_{i=1}^n P(i, j) \leq 1 \\ & P(i, j) = 0 \quad \text{or} \quad 1 \end{aligned}$$

Suppose $P(i, k_i) = 1$, then

$$\begin{aligned} \text{trace}\{P\Gamma\Gamma^TP^T\} &= \sum_{i=1}^n P(i, :) \Gamma \Gamma^T P^T(i, :) \\ &= \sum_{i=1}^n \{P(i, :) \Gamma\} \{P(i, :) \Gamma\}^T = \sum_{i=1}^n \sum_{j=1}^N \{\Gamma(k_i, j)\}^2 \end{aligned} \quad (6)$$

Define the sparsity preserving score of the i -th feature as,

$$\text{Score}(i) = \sum_{j=1}^N (\{\Gamma(k_i, j)\})^2 \quad (7)$$

Note that $\text{Score}(i) \geq 0$ for any i , and P can be simply solved by selected the n smallest ones from $\text{Score}(i), i = 1, \dots, m$. Without loss of generality, suppose the n selected features are indexed by $k_i^*, i = 1, 2, \dots, n$. We can construct the matrix P as

$$P(i, j) = \begin{cases} 1, & j = k_i^* \\ 0, & \text{otherwise} \end{cases}$$

Since the P obtained via the first iteration is 0-1 matrix, some values of features (corresponding to $j \neq k_i^*$) are equal to zero in the second iteration. Thus it is meaningless to compute the coefficient vector β_i for features whose values are equal to zero. In other words, P becomes a stable value after the first iteration. Thus, we simply solve the (3) in two steps.

The procedure of SPS is give in Algorithm 1.

- 1: **Input:** Training data $x_i (i = 1, \dots, N)$, λ , the number of feature selected n
- 2: Step 1: Solve $\beta_i^* : \beta_i^* = \arg \min_{\beta_i} \sum_{i=1}^N \|x_i - D_i\beta_i\|_F^2 + \lambda \|\beta_i\|_1$
- 3: Step 2: Solve P : compute $\text{Score}(i), i = 1, \dots, m$ and select the n smallest ones indexed by $k_i^*, i = 1, 2, \dots, n$. Let

$$P^*(i, j) = \begin{cases} 1, & j = k_i^* \\ 0, & \text{otherwise} \end{cases}$$
- 4: **Output:** solution $\beta_i^*, i = 1, \dots, N$ and P^*

Algorithm 1. Procedure for Sparsity Preserving Score

4 Experiments and Analysis

Several experiments on Yale and ORL face datasets are carried out to demonstrate the efficiency and effectiveness of our algorithm. Our algorithm is an unsupervised method, and thus we compared our algorithm with other four representative unsupervised feature selection algorithms including Data Variance, Laplacian Score, feature selection for multi-cluster data (MCFS) [17], and spectral feature selection (SPEC) [16] with all the eigenvectors of the graph Laplacian. In all the tests, the number of nearest neighbors in Laplacian Score, MCFS and SPEC is taken to be half of the number of training images per person.

For both datasets, we choose the first 5 and 6 images, respectively, per person for training and the rest for testing. After feature selection, the recognition is performed by the 'L2'-distance based 1-Nearest Neighbor classifier. Table 1 reports the top performance as well as the corresponding number of features selected, and Figure 1 illustrates the recognition rate as a function of the number of features selected. As shown in Table 1, our algorithm reaches the highest or comparable recognition rate at the lowest dimension of feature selected space. From Figure 1, we can see that with only a very small number of features, SPS can achieve significant better recognition rates than the other methods. It can be interpreted from two aspects: (1) SPS jointly selected features and obtain the optimal solution of a binary transformation matrix, while the other methods only add features one by one. Thus SPS considers the interaction and dependency among features. (2) Features selected with sparse reconstructive relationship preserving are capable of enhancing recognition performance.

Table 1. The comparison of the top recognition rates and the corresponding number of features selected

Methods	Training Date		Yale		ORL	
	5	6	5	6	5	6
Data Variance	0.6889(704)	0.6800(829)	0.9450(2503)	0.9563(2112)		
Laplacian Score	0.7111(434)	0.7067(952)	0.9450(2390)	0.9563(1901)		
MCFS	0.6556(974)	0.6933(825)	0.9250(1593)	0.9500(588)		
SPEC	0.7111(836)	0.7200(780)	0.9150 (2563)	0.9500(2350)		
SPS	0.7333(551)	0.7333(569)	0.9450(2355)	0.9563(1823)		

We randomly choose 5,6,7 and 8 images, respectively, per person for training and the rest for testing. Since the training set is randomly chosen, we repeat this experiment 10 times and calculate the average result. The average top performances obtained are reported in Table 2. The results further verify that SPS can select more informative preserving feature subset.

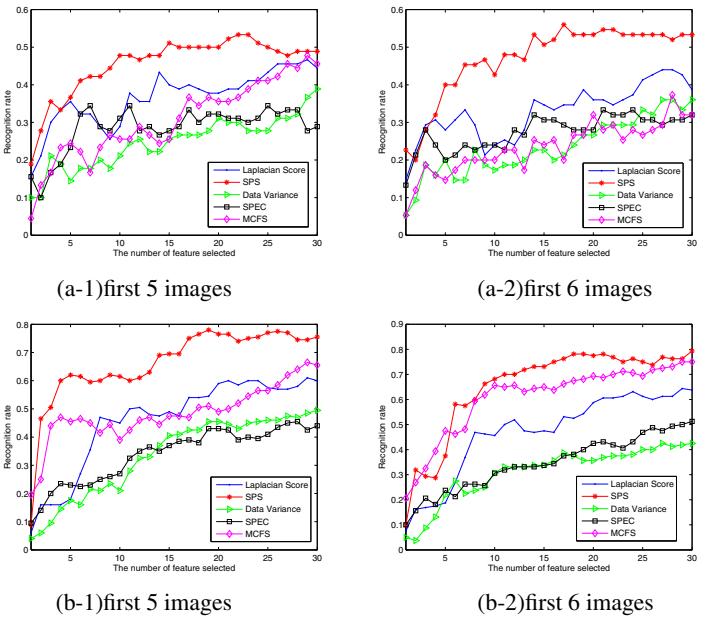


Fig. 1. Recognition results of the feature selection methods with respect to the number of selected features on (a) Yale and (b) ORL

Table 2. The comparison of average top recognition rates

(a) ORL		(b) Yale							
Methods	5	6	7	8	Methods	5	6	7	8
Data Variance	0.970	0.978	0.989	0.980	Data Variance	0.636	0.706	0.790	0.717
Laplacian Score	0.960	0.976	0.981	0.984	Laplacian Score	0.646	0.712	0.789	0.683
MCFS	0.950	0.958	0.960	0.955	MCFS	0.602	0.684	0.783	0.745
SPEC	0.940	0.947	0.958	0.950	SPEC	0.621	0.685	0.762	0.735
SPS	0.985	0.989	0.993	0.991	SPS	0.669	0.728	0.808	0.756

Acknowledgements. The authors would like to thank the anonymous reviewers for their constructive advice. This work is supported by the National Natural Science Foundation of China (Grant No.61202134), Jiangsu Planned Projects for Postdoctoral Research Funds, China Planned Projects for Postdoctoral Research Funds and National Science Fund for Distinguished Young Scholars (Grant No. 61125305).

References

1. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3, 1157–1182 (2003)
2. Gu, Q.Q., Li, Z.H., Han, J.W.: Joint Feature Selection and Subspace Learning. In: International Joint Conference on Artificial Intelligence, Barcelona, Spain (2011)

3. Nie, F.P., Huang, H., Cai, X., Ding, C.: Efficient and Robust Feature Selection via Joint $l_{2,1}$ -Norms Minimization. In: Advances in Neural Information Processing Systems, Vancouver, BC, Canada, pp. 1813–1182 (2010)
4. Ma, Z.G., Nie, F.P., Yang, Y., Uijlings, J.R.R., Sebe, N.: Web Image Annotation Via Subspace-Sparsity Collaborated Feature Selection. IEEE Transaction on Multimedia 14(4), 1021–1030 (2012)
5. Yan, H., Yuan, X.T., Yan, S.C., Yang, J.Y.: Correntropy based Feature Selection using Binary Projection. Pattern Recognition 44, 2834–2842 (2011)
6. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley Sons, NewYork (2001)
7. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust Face Recognition via Sparse Representation. IEEE Journal on Pattern Analysis and Machine Intelligence 31, 210–227 (2009)
8. Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: A Method for Largescale l_1 -regularized Least Squares. IEEE Journal on Selected Topics in Signal Processing 1(4), 606–617 (2007)
9. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artificial Intelligence 92(12), 273–324 (1997)
10. Devijver, P.A.: Pattern Recognition: A Statistical Approach. Prentice-Hall (1982)
11. Jain, A., Zongker, D.: Feature Selection: Evaluation, Application, and Small sample performance. IEEE Journal on Pattern Analysis and Machine Intelligence 19, 153–158 (1997)
12. Qiao, L.S., Chen, S.C., Tan, X.Y.: Sparsity Preserving Projections with Applications to Face Recognition. Pattern Recognition 43(1), 331–341 (2010)
13. Clemmensen, L., Hastie, T., Witten, D., Ersboll, B.: Sparse Discriminant Analysis. Technometrics 53(4), 406–413 (2011)
14. Zhang, L.M., Chen, S., Qiao, L.: Graph Optimization for Dimensionality Reduction with Sparsity Constraints. Pattern Recognition 45(3), 1205–1210 (2012)
15. He, X., Cai, D., Niyogi, P.: Laplacian Score for Feature Selection. In: Advances in Neural Information Processing Systems, Cambridge, MA (2005)
16. Zhao, Z., Liu, H.: Spectral Feature Selection for Supervised and Unsupervised Learning. In: Proceedings of International Conference on Machine Learning (ICML). ACM, New York (2007)
17. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)

Adaptive Backstepping Controller Design for Reentry Attitude of Near Space Hypersonic Vehicle

Jingmei Zhang¹, Changyin Sun¹, Ruimin Zhang¹, Chengshan Qian², and Lei Xue¹

¹ Flight Control Research Center, School of Automation, Southeast University,
Nanjing 210096, China

² College of Information and Control,
Nanjing University of Information Science & Technology, Nanjing 210044, China

Abstract. Based on backstepping design, a strongly robust adaptive control system is designed for reentry attitude of near space hypersonic vehicle, which is suffered by parameter uncertainties and external disturbances. In each step of backstepping design, the unknown upper bound of uncertainty is estimated by adaptive method to reduce conservatism. The derivative of the virtual controller is calculated by a precise differentiator based on higher-order sliding modes. Using non-quadratic Lyapunov function, the global asymptotic stability of the closed-loop system is proved. Simulation results show that the proposed control system can overcome the impact of large-scale perturbations of aerodynamic parameters and external disturbances, which has better dynamic qualities, tracking capabilities compared to the traditional backstepping controller.

Keywords: Hypersonic vehicle, backstepping design, adaptive sliding mode controller gain, attitude control.

1 Introduction

During the re-entry attitude of near space hypersonic vehicle, because of the complex flight environment, wide range of the speed and the flight height, and the tremendous changes of gas thermal property and aerodynamic characteristics, the basic structure of the control system is lack of fidelity and the models or parameters become uncertain [1]. Under the condition of hypersonic flight, the hypersonic vehicle will be affected by the hydrodynamic effect, effect of viscosity and rarefied gas effect. Therefore, the hypersonic vehicle is a complex nonlinear system with much uncertainties. Moreover, the uncertainties in its dynamics model [2] do not meet the parameter matching conditions, so it requires the designed control system should have strong robustness and adaptive ability.

Backstepping design method [3] is an effective design method for nonlinear control system. It shows unique advantages in dealing with nonlinear system problems. Take cascade linear or nonlinear system for example, by selecting the suitable Lyapunov function and constructing virtual control laws back step by step, a stable control system can be got. This method can guarantee the global stability of closed-loop system, regulating and tracking has asymptotic behavior. But the

traditional adaptive backstepping design method acquires the system can be parameterized [4~6]. Reference [7] proposed a robust backstepping sliding mode control method for nonlinear systems by using the sliding mode control to compensate the influence of the uncertainty. However, the sliding-mode controller is designed based on the upper bound of the uncertainty, it acquire that all the uncertainties of sub systems should be known. Reference [8] combined the backstepping control, adaptive control, and sliding-mode control together and proposed a new adaptive backstepping controller. However, the derivation of virtual control mentioned in this paper is treated as uncertainty, so this approach brings the conservative to design the controller.

This paper proposes a control method which combines the adaptive control and backstepping design together. This approach can overcome the limitation of acquirement that the uncertainty can be parameterized, as well as enhances the robustness of mismatched uncertainties. Therefore, this paper provides a valid approach to attitude tracking control of near space hypersonic vehicle.

2 Problem Description

The states of system are $x_1 = [\alpha, \beta, \infty]^T$, $x_2 = [p, q, r]^T$. The input of the system is $u = g_{f\delta}\delta + M_r$. Therefore the re-entry attitude of the near space hypersonic vehicle can be described as follows[2]:

$$\dot{x}_1 = f_s(x_s) + G_{s1}(x_s)x_2 + \Phi_1 \quad (1)$$

$$\dot{x}_2 = f_f(x_f) + G_f(x_f)u + \Phi_2 \quad (2)$$

where, Φ_1, Φ_2 are compound disturbance which contain the uncertainty caused by pneumatic parameter perturbation and the external disturbance.

$$\begin{aligned} f_s(x_s) &= [f_\alpha, f_\beta, f_\infty]^T, f_\alpha = (-\hat{q}SC_{L\alpha} + Mg \cos \gamma \cos \alpha) / (MV \cos \beta), \\ f_\beta &= (\hat{q}SC_{Y\beta} \beta \cos \beta + Mg \cos \gamma \sin \alpha) / (MV), \\ f_\infty &= -g \cos \gamma \cos \alpha \tan \beta / V + \hat{q}SC_{Y\beta} \beta \tan \gamma \cos \alpha \cos \beta / (MV) + \hat{q}SC_{L\alpha} (\tan \gamma \sin \alpha + \tan \beta) / (MV). \end{aligned}$$

$$G_{s1}(x_s) = \begin{bmatrix} -\tan \beta \cos \alpha & 1 & -\tan \beta \sin \alpha \\ \sin \alpha & 0 & -\cos \alpha \\ \sec \beta \cos \alpha & 0 & \sec \beta \sin \alpha \end{bmatrix}, \quad \begin{aligned} f_f(x_s) &= [f_p, f_q, f_r]^T, \\ f_p &= I_{qr}^p qr + I_p^p p + g_l^p l_{aero}, \\ f_q &= I_{pr}^q pr + I_q^q q + g_m^q m_{aero}, \end{aligned}$$

$$\begin{aligned} f_r &= I_{pq}^r pq + I_r^r r + g_n^r n_{aero}, I_{qr}^p = (I_{yy} - I_{zz}) / I_{xx}, I_p^p = -g_l^p \dot{I}_{xx}, g_l^p = 1 / I_{xx}, \dot{I}_q^q = -g_m^q \dot{I}_{yy}, \\ I_{pr}^q &= (I_{zz} - I_{xx}) / I_{yy}, g_m^q = 1 / I_{yy}, I_{pq}^r = (I_{xx} - I_{yy}) / I_{zz}, \dot{I}_r^r = -g_n^r \dot{I}_{zz}, g_n^r = 1 / I_{zz}, \\ l_{aero} &= \hat{q}Sb(C_{l,\beta}\beta + C_{l,p}pb/2V + C_{l,r}rb/2V), m_{aero} = \hat{q}Sc\left(C_{m,\alpha} + C_{m,q}\frac{qc}{2V}\right) + X_{cg}\hat{q}SC_{Y\beta}\beta, G_f = diag(g_l^p, g_m^q, g_n^r), \\ n_{aero} &= \hat{q}Sb(C_{n,\beta}\beta + C_{n,p}pb/2V + C_{n,r}rb/2V) + X_{cg}\hat{q}SC_{Y\beta}\beta, G_f = diag(g_l^p, g_m^q, g_n^r), \end{aligned}$$

where, $\delta = [\delta_e, \delta_a, \delta_r]^T$ are the angles of both sides elevons and rudder deflection; $M_r = [l_{Tr}, m_{Tr}, n_{Tr}]^T$ is the decomposition for moment of thrust of reaction control system in body coordinates;

$$g_{f\delta} = \begin{bmatrix} g_{p,\delta_e} & g_{p,\delta_a} & g_{p,\delta_r} \\ g_{q,\delta_e} & g_{q,\delta_a} & g_{q,\delta_r} \\ g_{r,\delta_e} & g_{r,\delta_a} & g_{r,\delta_r} \end{bmatrix}, \quad \begin{aligned} g_{p,\delta_e} &= \hat{q}SbC_{l,\delta_e}, \quad g_{p,\delta_a} = \hat{q}SbC_{l,\delta_a}, \quad g_{p,\delta_r} = \hat{q}SbC_{l,\delta_r}, \\ g_{q,\delta_e} &= \hat{q}ScC_{m,\delta_e} + X_{cg}\hat{q}S(C_{D,\delta_e}\sin\alpha + C_{L,\delta_e}\cos\alpha), \\ g_{q,\delta_a} &= \hat{q}ScC_{m,\delta_a} + X_{cg}\hat{q}S(C_{D,\delta_a}\sin\alpha + C_{L,\delta_a}\cos\alpha), \\ g_{r,\delta_e} &= \hat{q}SC_{m,\delta_e} + X_{cg}\bar{q}SC_{D,\delta_e}\sin\alpha, \quad g_{r,\delta_a} = \hat{q}SbC_{n,\delta_e} + X_{cg}\hat{q}SC_{Y,\delta_e}, \\ g_{r,\delta_a} &= \hat{q}SbC_{n,\delta_a} + X_{cg}\hat{q}SC_{Y,\delta_a}, \quad g_{r,\delta_r} = \hat{q}SbC_{n,\delta_r} + X_{cg}\hat{q}SC_{Y,\delta_r}. \end{aligned}$$

Assumption 1. The gain matrices of control system G_{s1} , G_f can be invertible.

Assumption 2. The uncertainties of NSV satisfy $\|\Phi_i\| \leq \varphi_i$. Where the φ_i is unknown bounded normal number $i = 1, 2$.

The main purpose of this paper is that designing the robust adaptive virtual control law x_{2d} and control law u based on backstepping method. In this way, the re-entry attitude of NSV x_1 can track the given target x_{1d} asymptotically.

3 The Robust Adaptive Backstepping Control Design

First step: Designing the virtual control law to compensate the influence of uncertainty at the attitude angle loop (1).

Define the error state vectors:

$$z_1 = x_1 - x_{1d} \quad (3)$$

$$z_2 = x_2 - x_{2d} \quad (4)$$

Due to the equation (1) and (3), the dynamic equation can be described as:

$$\dot{z}_1 = f_s(x_s) + G_{s1}(x_s)x_2 + \Phi_1 - \dot{x}_{1d} \quad (5)$$

Because of the Assumption 1, the virtual control matrix $G_{s1}(x_s)$ can be invertible. The virtual control law is designed as follows:

$$x_{2d} = G_{s1}^{-1}(x_s)[-f_s(x_s) - \hat{\varphi}_1^2 z_1 / (\hat{\varphi}_1 \|z_1\| + \varepsilon_1 e^{-a_1 t}) - k_1 z_1 + \dot{x}_{1d}] \quad (6)$$

where $\varepsilon_1 > 0$, $a_1 > 0$, $\hat{\varphi}_1$ is the estimated value of φ_1 . The adaptive law is shown as follows:

$$\dot{\hat{\varphi}}_1 = \gamma_1 \|z_1\|, \quad \gamma_1 > 0 \quad (7)$$

Considering the following Lyapunov function:

$$V_1 = 1/2 z_1^T z_1 + 1/2 \gamma_1 \tilde{\varphi}_1^2 + \varepsilon_1 / a_1 e^{-a_1 t}$$

The time derivative is:

$$\begin{aligned}\dot{V}_1 &= z_1^T \dot{z}_1 + 1/\gamma_1 \cdot \tilde{\varphi}_1 \dot{\tilde{\varphi}}_1 - \varepsilon_1 e^{-a_1 t} = z_1^T (f_s + G_{s1} z_2 + G_{s1} x_{2d} + \Phi_1 - \dot{x}_{1d}) + 1/\gamma_1 \cdot \tilde{\varphi}_1 \dot{\tilde{\varphi}}_1 - \varepsilon_1 e^{-a_1 t} \\ &= z_1^T (G_{s1} z_2 - k_1 z_1 + \Phi_1 - \hat{\varphi}_1^2 z_1 / (\hat{\varphi}_1 \|z_1\| + \varepsilon_1 e^{-a_1 t})) + \hat{\varphi}_1 \|z_1\| - \varphi_1 \|z_1\| - \varepsilon_1 e^{-a_1 t} \\ &\leq z_1^T (G_{s1} z_2 - k_1 z_1 - \hat{\varphi}_1^2 z_1 / (\hat{\varphi}_1 \|z_1\| + \varepsilon_1 e^{-a_1 t})) + \hat{\varphi}_1 \|z_1\| - \varepsilon_1 e^{-a_1 t} \leq z_1^T G_{s1} z_2 - z_1^T k_1 z_1\end{aligned}$$

Second step: Designing the control law u to make the global tracking errors z_1 and z_2 converge to zero asymptotically.

$$\dot{z}_2 = \dot{x}_2 - \dot{x}_{2d} = f_f + G_f u + \Phi_2 - \dot{x}_{2d}, \quad (8)$$

The control law u is designed as follows:

$$u = -G_f^{-1} [f_f + k_2 z_2 + \hat{\varphi}_2^2 z_2 / (\hat{\varphi}_2 \|z_2\| + \varepsilon_2 e^{-a_2 t}) + G_{s2}^T z_1 - \dot{x}_{2d}] \quad (9)$$

where, $\varepsilon_2 > 0$, $a_2 > 0$, $\hat{\varphi}_2$ is the estimated value of φ_2 . The adaptive law is:

$$\dot{\hat{\varphi}}_2 = \gamma_2 \|z_2\|, \quad \gamma_2 > 0 \quad (10)$$

Considering the following Lyapunov function:

$$V_2 = 1/2 z_2^T z_2 + 1/2 \gamma_2 \hat{\varphi}_2^2 + \varepsilon_2 / a_2 e^{-a_2 t}$$

The time derivative is:

$$\begin{aligned}\dot{V}_2 &= z_2^T \dot{z}_2 + 1/\gamma_2 \tilde{\varphi}_2 \dot{\hat{\varphi}}_2 - \varepsilon_2 e^{-a_2 t} \\ &= z_2^T (f_f + G_f u + \Phi_2 - \dot{x}_{2d}) + 1/\gamma_2 \tilde{\varphi}_2 \dot{\hat{\varphi}}_2 - \varepsilon_2 e^{-a_2 t} \\ &= z_2^T (-k_2 z_2 - \hat{\varphi}_2^2 z_2 / (\hat{\varphi}_2 \|z_2\| + \varepsilon_2 e^{-a_2 t}) + \Phi_2 - G_{s2}^T z_1) + (\hat{\varphi}_2 - \varphi_2) \|z_2\| - \varepsilon_2 e^{-a_2 t} \\ &\leq -z_2^T k_2 z_2 - z_2^T G_{s2}^T z_1 - \hat{\varphi}_2^2 \|z_2\|^2 / (\hat{\varphi}_2 \|z_2\| + \varepsilon_2 e^{-a_2 t}) + \hat{\varphi}_2 \|z_2\| - \varepsilon_2 e^{-a_2 t} \\ &= -z_2^T k_2 z_2 - z_2^T G_{s2}^T z_1\end{aligned}$$

Therefore, according to Lyapunov stability theorem, the augmented errors $\Xi = [z_1^T \quad \tilde{\varphi}_1 \quad z_2^T \quad \tilde{\varphi}_2]^T$ of closed loop system are global asymptotically stable.

To sum up, robust adaptive control system (6),(7), (9),(10) can make the re-entry attitude of NSV x_i asymptotic tracking the given target x_{id} .

Remark 1. It is very complicated to calculate the $\dot{x}_{2d} = [\dot{x}_{2d1} \quad \dot{x}_{2d2} \quad \dot{x}_{2d3}]^T$ based on the equation (6). In this paper, the robust differentiator^[9] is mentioned to handle this problem.

$$\dot{z}_{0i} = -\lambda_{0i} |z_{0i} - x_{2di}|^{1/2} \operatorname{sign}(z_{0i} - x_{2di}) + z_{1i}, \quad \dot{z}_{1i} = -\lambda_{1i} \operatorname{sign}(z_{1i} - \alpha_i), \quad i = 1, 2, 3$$

where, $\lambda_{1i} > 0$, $\lambda_{0i} > 0$, Levant proved that when there is no noise, the robust differentiator can estimate the first-order derivative of the input signal. That is to say, the $z_{0i} = \alpha_i$, $z_{1i} = \dot{\alpha}_i$ was ensured within the finite time. When there is some noise, the estimate error can be offset by the robustness items.

4 Simulation

The initial conditions of simulation are $V_0 = 3.1 \text{ km/s}$, $H_0 = 34 \text{ km}$, $\alpha_0 = 0.1^\circ$, $\beta_0 = 0^\circ$, $\gamma_0 = 0^\circ$, $p_0 = q_0 = r_0 = 0^\circ/\text{s}$. Command reference signals are $\alpha_c = 1.5^\circ$, $\beta_c = 0^\circ$, $\gamma_c = 0.8^\circ$, given by the command filter $2/(s+2)$. It supposes the aerodynamic parameter has 50% uncertainty. At the same time, the disturbance torques were added into three Channels respectively, $10^5 \sin(t) N\cdot m$, $10^5 \cos(2t) N\cdot m$, $10^5 \sin(3t) N\cdot m$.

The parameters of the control system are given by $k_1 = \text{diag}\{1.2 \ 1.2 \ 1.2\}$, $k_2 = \text{diag}\{2 \ 2 \ 2\}$, $a_1 = 2$, $\varepsilon_1 = 0.5$, $a_2 = 1$, $\varepsilon_2 = 0.5$, $\gamma_1 = 0.8$, $\gamma_2 = 20$.

In order to illustrate the effectiveness of the method mentioned in this paper, it is compared with the traditional Backstepping method. According to the traditional Backstepping method, the control system of The NSV is shown as follows:

$$\bar{x}_2 = G_{s1}^{-1}(x_s)[-f_s(x_s) - k_1 z_1 + \dot{x}_{1d}], u = -G_f^{-1}[f_f + k_2 z_2 + G_{s1}^T z_1 - \dot{\bar{x}}_2]$$

The parameters of the system are $k_1 = \text{diag}\{15 \ 15 \ 15\}$, $k_2 = \text{diag}\{20 \ 20 \ 20\}$.

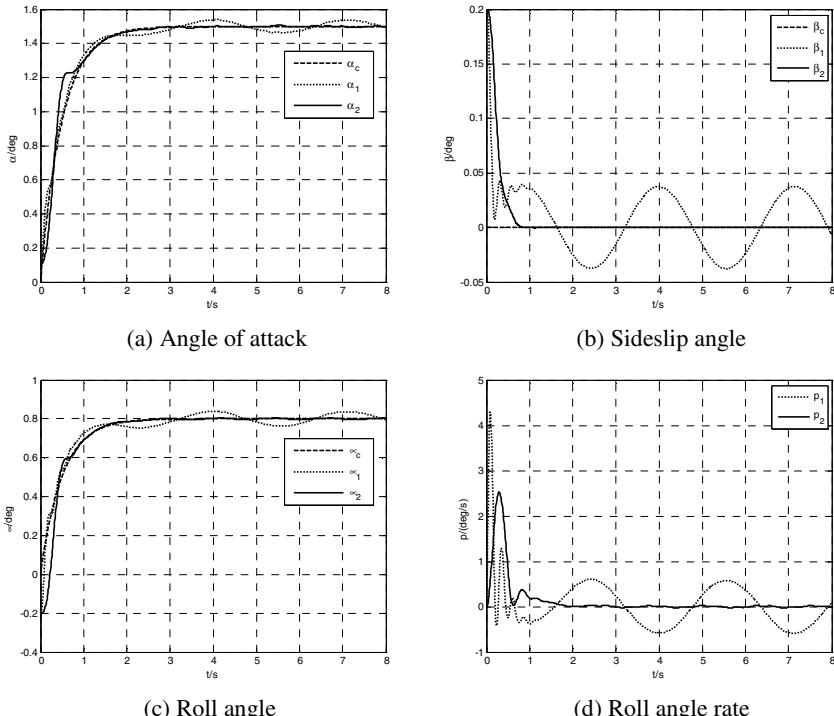
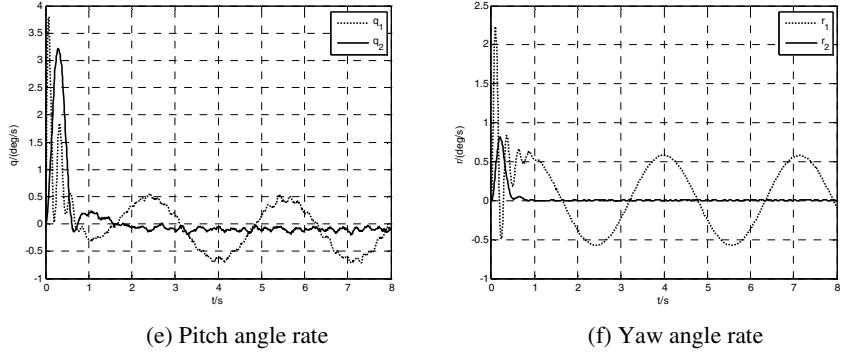


Fig. 1. Attitude angle command and angular rate tracking curves comparison



(e) Pitch angle rate

(f) Yaw angle rate

Fig. 1. (Continued.)

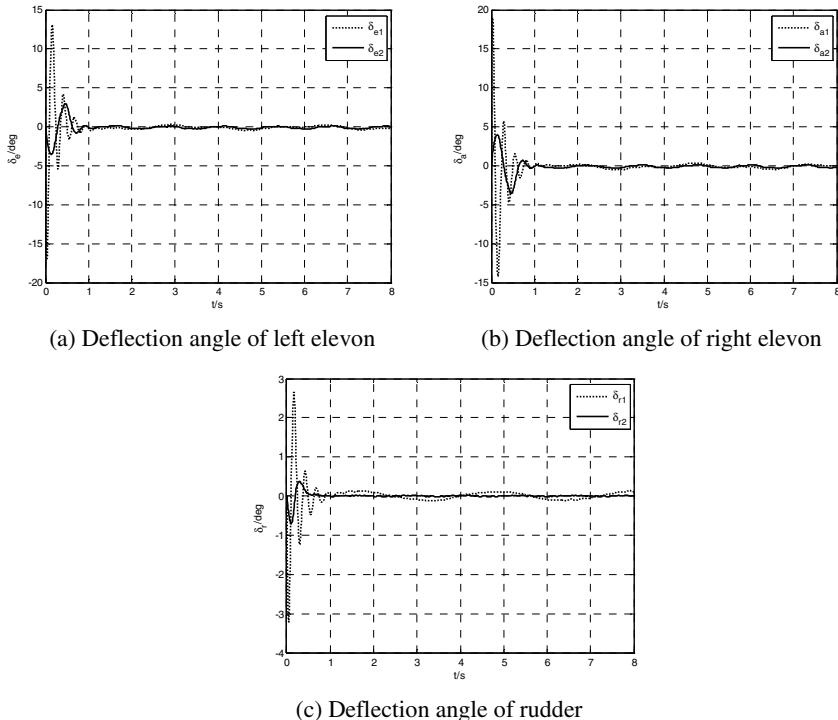


Fig. 2. The aero-surface control outputs curves comparison

The simulation results are shown as Fig.1 and Fig.2. Where subscript c means command signal, subscript 1, 2 express the simulation result of traditional backstepping control method and method proposed in this paper. Fig 1 shows the changing curve of attitude angle and angular rate. It is observed that the performance of robust adaptive backstepping control method is much better than the one of traditional backstepping control method. Fig. 2 shows the changing curve of the

steering engines. Based on the traditional backstepping control of rudder surface deflection, because of the need to adopt high control gain to suppress parameter uncertainty and external disturbance, results in the higher control efforts.

5 Conclusion

Concerning with reentry attitude tracking control of near space hypersonic vehicle, a strongly robust adaptive control system is proposed based on backstepping design and adaptive control technique. Specially, a robust adaptive virtual control law is designed by using the adaptive method to estimate the unknown upper bound of uncertainties both in the attitude angle loop and angular velocity loop. Using Lyapunov stability theorem, the tracking errors are shown to be asymptotic stability. Simulation results show that the proposed control system can overcome the impact of large-scale perturbations of aerodynamic parameters and external disturbances, which has better dynamic qualities, tracking capabilities compared to the traditional backstepping control system.

Acknowledgement. This work is supported by National Outstanding Youth Science Foundation (61125306); National Natural Science Foundation of Major Research Plan (91016004, 61034002), Specialized Research Fund for the Doctoral Program of Higher Education of China (20110092110020).

References

1. Jie, H., Dezhai, Z.: Characteristics and key technology research for hypersonic vehicle. In: Aerospace Science and Technology Innovation and the Yangtze River Delta Economic Transformation Development BBS, pp. 227–231 (2012)
2. Shaughnessy, J.D., Pinckney, S.Z., McMinn, J.D., et al.: Hypersonic vehicle simulation model: winged-cone configuration. NASA TM-102610 (1990)
3. Kanellakopoulos, I., Kokotovic, P.V., Stephen Morse, A.: Systematic design of adaptive controllers for feedback linearizable systems. IEEE Transactions on Automatic Control 36(11), 1241–1253 (1991)
4. Koshkouei, A.J., Zinober, A.S.I.: Adaptive backstepping control of nonlinear systems with unmatched uncertainty. In: Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, pp. 4765–4770 (December 2000)
5. Swaroop, D., Hedrick, J.K., Yip, P.P., Gerdes, J.C.: Dynamic surface control for a class of nonlinear systems. IEEE Transactions on Automatic Control 45(10), 1893–1899 (2000)
6. Gao, D., Sun, Z., Du, T.: Dynamic surface control for hypersonic aircraft using fuzzy logic system. In: Proceeding of IEEE International Conference on Automation and Logistics, Jinan, pp. 2314–2319 (2007)
7. Zhou, Y., Wu, Y., Hu, Y.: Robust backstepping sliding mode control of a class of uncertain MIMO nonlinear systems. In: 2007 IEEE International Conference on Control and Automation, Guangzhou, pp. 1916–1921 (2007)
8. Zhu, K., Qi, N., Qin, C.: Adaptive sliding mode controller design for BTT missile based on backstepping control. Journal of Astronautics 31(3), 769–773 (2011)
9. Levant, A.: Higher-order sliding modes, differentiation and output-feedback control. International Journal of Control 76(9/10), 924–941 (2003)

High Performance Super-Resolution Reconstruction of Multiple Images Based on Fast Registration and Edge Enhancement

Ming Liu¹, Jianyu Huang², Ming Gao¹, and Shiyin Qin¹

¹ School of Automation Science and Electrical Engineering, Beihang University,
100191, Beijing, China

² Beijing Institute of Tracking and Telecommunications Technology,
100094, Beijing, China
maxence.liu@gmail.com, jianyu2875@126.com,
minggao818@163.com, qsy@buaa.edu.cn

Abstract. In this paper, an approach to super-resolution (SR) reconstruction of multi-images is proposed based on improved Keren registration and a revised regularization method, which is characterized with high performance of edge enhancement and processing efficiency. In order to increase the registration speed, the Keren registration method is improved by partition of original images and parallel registration of each small patch. And then a revised regularization method is employed to resolve the ill-posed problem and perform the edge enhancement simultaneously. The experimental results indicate that the processing efficiency is raised in a large extent, provided the image registration precision is as high as the classical methods. In the SR reconstruction effect, it can outperform other super-resolution methods observably, thus takes more advantages in practical applications.

Keywords: image registration, super-resolution, edge enhancement, regularization.

1 Introduction

It is always attractive and challenging to obtain high-resolution (HR) images. The main purpose of super-resolution (SR) algorithm is to utilize a set of low-resolution (LR) aliased images to reconstruct one or a set of HR images. Before solving this inverse problem [1][2], the motion information between images needs to be estimated by image registration. There are already several methods to resolve this problem. Such as the method proposed by Keren [3] and Optical Flow method [4] proposed by Horn and Schunck. These two methods can both deal with the sub-pixel motion estimation, and have very high estimation precision. Another motion estimation method is block motion estimation [5], but its accuracy is relatively low. The precision of motion estimation and the computational complexity of the algorithm will directly influence the latter SR algorithm's effect and real-time capacity.

Various types of SR algorithm have been proposed during the last two decades. The multi-image SR algorithm was first proposed by Huang and Tsai [8], where they dealt

with the problem in the frequency domain. After that, solving the SR reconstruction problem in the spatial domain using multi-image becomes a tendency. The non-uniform interpolation-based methods [9] have very low computational cost, but they do not take into account the blur model and noise characteristics. The Projection onto Convex Set (POCS) [10] based methods combine the spatial domain observation model and prior information. But they have the slow convergence rate and the high computational cost. Maximum-likelihood (ML) [11] and maximum a posteriori (MAP) [12] are two widely used SR reconstruction algorithms. But they do not perform very well in preserving the edge in the initial stage. Within the framework of MAP, Farsiu proposed the bilateral total variation (BTV) regularization in [13], which has successfully improved the denoising and deblurring performance with edge preserving. The learning based algorithm [14][15] is another class of SR reconstruction methods, these methods generate a HR image from single frame or image.

In this paper, we firstly introduce an improved Keren based motion estimation algorithm in Section 2. Then, a new regularization method for SR reconstruction is proposed in Section 3, which is proved to be able to remove image noise and enhance edge information.

2 Improved Image Registration towards SR Reconstruction

In this section, we firstly assume that the motion between images is pure translation and rotation, and the extent of translation and rotation is not too big. Although these preconditions are limiting to images, they conform to real situation of the SR reconstruction.

We use the following approach to calculate the motion parameters. Function $f(x, y)$ and $g(x, y)$ are used to represent the two images. Then we divide them into several identical sub-images $\{f_k\}_{k=1}^N$ and $\{g_k\}_{k=1}^N$. Fig.1 shows the way to create sub-images from an original image.

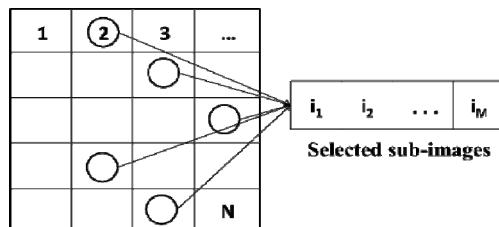


Fig. 1. Strategy of the creation of sub-images

There are altogether N sub-images. Then we select a subset of $\{f_k\}_{k=1}^N$ and $\{g_k\}_{k=1}^N$ to do the image registration. For each pair of sub-images f_k and g_k ($k \in \{i_l\}_{l=1}^M$), we use vertical shift parameter a_k and horizontal shift parameter b_k to represent the relative motion vector. Then the relationship between two sub-images can be expressed as:

$$g_k(x, y) = f_k(x + a_k, y + b_k), \quad k \in \{i_l\}_{l=1}^M \quad (1)$$

where f_k can be expanded to its Taylor series with its first order equation:

$$g_k(x, y) = f_k(x, y) + a_k \frac{\partial f_k}{\partial x} + b_k \frac{\partial f_k}{\partial y} \quad (2)$$

Then the error function between f_k and g_k can be approximately expressed as:

$$E(a_k, b_k) = \sum_{\Omega_k} [f_k(x, y) + a_k \frac{\partial f_k}{\partial x} + b_k \frac{\partial f_k}{\partial y} - g_k(x, y)]^2 \quad (3)$$

The summation in (3) is over the overlapping part Ω_k of the sub-images f_k and g_k . Then we look for the minimum of $E(a_k, b_k)$ by computing its derivatives by a_k and b_k , and obtain the following system of linear equations:

$$\begin{aligned} a_k \sum_{\Omega_k} \left(\frac{\partial f_k}{\partial x} \right)^2 + b_k \sum_{\Omega_k} \frac{\partial f_k}{\partial x} \frac{\partial f_k}{\partial y} &= \sum_{\Omega_k} (g_k - f_k) \frac{\partial f_k}{\partial x} \\ a_k \sum_{\Omega_k} \frac{\partial f_k}{\partial x} \frac{\partial f_k}{\partial y} + b_k \sum_{\Omega_k} \left(\frac{\partial f_k}{\partial y} \right)^2 &= \sum_{\Omega_k} (g_k - f_k) \frac{\partial f_k}{\partial y} \end{aligned} \quad (4)$$

In order to get a simple expression, we note

$$Z_k = \begin{pmatrix} \sum_{\Omega_k} (g_k - f_k) \frac{\partial f_k}{\partial x} \\ \sum_{\Omega_k} (g_k - f_k) \frac{\partial f_k}{\partial y} \end{pmatrix}, C_k = \begin{pmatrix} \sum_{\Omega_k} \left(\frac{\partial f_k}{\partial x} \right)^2 & \sum_{\Omega_k} \frac{\partial f_k}{\partial x} \frac{\partial f_k}{\partial y} \\ \sum_{\Omega_k} \frac{\partial f_k}{\partial x} \frac{\partial f_k}{\partial y} & \sum_{\Omega_k} \left(\frac{\partial f_k}{\partial y} \right)^2 \end{pmatrix} \text{ and } R_k = \begin{pmatrix} a_k \\ b_k \end{pmatrix} \quad (5)$$

then (4) can be reformulated as $C_k R_k = Z_k$. In order to augment the accuracy of the motion estimation, we perform the following iterations:

$$R_k^{n+1} = R_k^n + C_k^{-1} Z_k \quad (6)$$

Within each iteration, we do the sub-pixel interpolation for the original g_k by using the values of R_k^n . The selected sub-images should better be representative for the image. Thus we can remarkably reduce the computational cost.

With the translation results $\{(a_k, b_k)\}$ ($k \in \{i_l\}_{l=1}^M$) of each sub-images, we can derive the translation parameters a and b and the rotation θ between image f and g . For each sub-image, we use the center coordinate (x_k, y_k) to represent its original position, and $(x_k + a_k, y_k + b_k)$ as the coordinate after the translation. According to the rigid motion assumption, we get the following expression:

$$\begin{pmatrix} \cos \theta & -\sin \theta & a \\ \sin \theta & \cos \theta & b \end{pmatrix} \begin{pmatrix} x_k \\ y_k \\ 1 \end{pmatrix} = \begin{pmatrix} x_k + a_k \\ y_k + b_k \end{pmatrix}, \quad k \in \{i_l\}_{l=1}^M \quad (7)$$

There are $2M$ equations and 4 variables, including $\cos \theta$, $\sin \theta$, a and b . Then (7) can be solved by the least square method. With the exact registration results of the image sequence, we can perform the subsequent SR image reconstruction process.

3 High Performance SR Reconstruction Based on Fast Registration and Edge Enhancement

3.1 Promotion of SR Reconstruction by High Precision Registration and Edge Enhancement

The degradation of the image sequence is usually due to the atmospheric turbulence, inappropriate camera settings, downsampling determined by the resolution capacity of camera and the noise produced by sensor. Most papers [6][7][13] describe the relationship between LR images and HR image as the following formula:

$$Y_k = DF_k H_k X + V_k, \quad \forall k = 1, 2, \dots, K \quad (8)$$

where $\{Y_k\}_{k=1}^K$ are the K captured LR images of size $MN \times 1$, X represents the HR image of size $r^2 MN \times 1$ where r is the resolution enhancement factor. V_k is the additive white Gaussian noise. F_k is the geometric warp matrix and H_k is the blurring matrix of size $r^2 MN \times r^2 MN$. D is the downsampling matrix of size $MN \times r^2 MN$. Since we assume that all the LR images are taken under the same circumstance and by the same sensor, H_k becomes the same for all k and can be simplified as H . Then the equation (9) can be expressed as:

$$Y_k = DF_k HX + V_k, \quad \forall k = 1, 2, \dots, K \quad (9)$$

Here we use MAP-based SR to solve this ill-posed problem. The desired HR image can be formulated [12][13][16] as:

$$\hat{X} = \arg \min_X \left\{ \sum_{k=1}^K \|DF_k HX - Y_k\|_2^2 + \frac{\alpha}{2} \|\Gamma * X\|_2^2 \right\} \quad (10)$$

The regularization parameter μ is set to adjust the proportion between the data error and the smoothness of image. One of the most widely used regularization method of Γ is Laplacian kernel. The operator $*$ is the convolution operator. Then we calculate the derivatives of (10) and use steepest descent (SD) algorithm to find its solution:

$$\hat{X}_{n+1} = \hat{X}_n + \beta \left[\sum_{k=1}^K (H^T F_k^T D^T (DF_k H \hat{X}_n - Y_k)) + \alpha \underline{\Gamma}' \underline{\Gamma} X \right] \quad (11)$$

where \hat{X}_0 is the initialization value of HR image. β is the step size in the direction of gradient.

If we use Laplacian kernel as the regularization method, it will remove the noise as well as the edge information. One of the successful edge-preserving regularization method which can also remove noise is total variation (TV) method[17]. Farsiu improved the TV method by bilateral total variation (BTV) which has better performance than TV. But the computational complexity is much more than high-pass regularization method.

Considering the disadvantage of Laplacian kernel, we introduce here a combinational operator which can remove the noise and the enhance edge at the same time. The regularization function looks like:

$$\gamma(X) = \frac{1}{2} \left[\left\| \Gamma^{0.5} X \right\|_2^2 + \alpha \left(\|S * X\|_2^2 + \|S' * X\|_2^2 \right) \right] \quad (12)$$

Since the matrix $\underline{\Gamma}$ is a block-circulant matrix, it is a unitary diagonalizable matrix which can be expressed as:

$$\underline{\Gamma}^{0.5} = F^* D^{0.5} F \quad (13)$$

where F is the modal matrix whose columns are the eigenvectors of $\underline{\Gamma}$. Matrix D is the diagonal matrix with the eigenvalues of $\underline{\Gamma}$. The index 0.5 in (13) means the element-wise square root of the diagonal matrix D . The kernel S is the Sobel operator and its transposition S' represent the vertical and horizontal edge enhancing operators respectively. According to (10), the SR becomes the following minimization problem:

$$\hat{X} = \arg \min_X \left\{ \sum_{k=1}^K \|DF_k HX - Y_k\|_2^2 + \frac{\alpha}{2} \left[\left\| \Gamma^{0.5} X \right\|_2^2 + \alpha \left(\|S * X\|_2^2 + \|S' * X\|_2^2 \right) \right] \right\} \quad (14)$$

To compare the performance of our method with the BTV and Tikhonov regularization method, we set up the following experiment. We degrade an image Fig.2 (a) to Fig.2 (b) with a Gaussian kernel of size 5×5 and add Gaussian additive white noise with the variance of 0.01. We reconstruct the image with Tikhonov, BTV and our proposed regularization method. Through modifying function (14), we minimize:

$$\hat{X} = \arg \min_X \left\{ \|X - Y\|_2^2 + \alpha \gamma(X) \right\} \quad (15)$$

to reconstruct the image. Tikhonov denoising result is shown in Fig.2 (c), the noise is removed to some extent, but it also brings severe loss of sharp edge information. The result of BTV is shown in Fig.2 (d), BTV can efficiently remove the noise, and its edge-preserving effect is obvious, but the blurred edge is also preserved by this approach. What's more, the two letters fade in gray level and the background becomes muddy. The result of our method is shown in Fig.2 (e). There is much less noise than the noisy image and the result of Tikhonov method. The gray level of the two letters and the background look much better than Fig.2 (d).

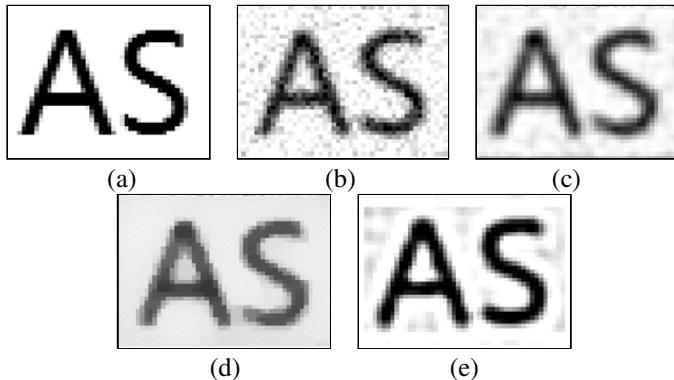


Fig. 2. Results of denoising (a) Original .(b) Noisy and blur. (c) Reconstruction using Tikhonov. (d) Reconstruction using BTV. (e) Reconstruction using edge enhancement method.

3.2 Implementation of High Performance SR Algorithm

With the ideas presented in Section 3.A, and the property of the matrix in the regularization term, we calculate the derivatives of the right part of function (14), and use the SD algorithm to find the solution of this minimization problem:

$$\begin{aligned}\hat{X}_{n+1} &= \hat{X}_n + \beta \left[\sum_{k=1}^K (H^T F_k^T D^T (DF_k H \hat{X}_n - Y_k)) + \alpha \gamma'(\hat{X}_n) \right] \\ &= \hat{X}_n + \beta \left[\sum_{k=1}^K (H^T F_k^T D^T (DF_k H \hat{X}_n - Y_k)) + \alpha (\underline{\Gamma}^{0.5} \underline{\Gamma}^{0.5} \hat{X}_n + \alpha (\underline{S}' \underline{S} + \underline{S}'' \underline{S}') \hat{X}_n) \right]\end{aligned}\quad (16)$$

where we use the diagonalized matrix (13) to simplify the (16):

$$\underline{\Gamma}^{0.5} \underline{\Gamma}^{0.5} = F^* \text{sqrt}(D^*) F F^* \text{sqrt}(D) F = F^* D F = \underline{\Gamma}\quad (17)$$

And due to $\underline{S}' = -\underline{S}$, the iterative process of the optimization becomes:

$$\hat{X}_{n+1} = \hat{X}_n + \beta \left[\sum_{k=1}^K (H^T F_k^T D^T (DF_k H \hat{X}_n - Y_k)) + \alpha (\underline{\Gamma} \hat{X}_n + \alpha (\underline{S} \underline{S} + \underline{S}' \underline{S}') \hat{X}_n) \right]\quad (18)$$

where α is the edge enhancing factor. The matrix \underline{S} and \underline{S}' define the block-circulant matrix of the sobel kernel and its transposition respectively.

4 Experiment Results

In this section, we perform two sets of experiments (with computer Core2 2.5GHz) to test the efficiency and the precision of our registration method and the reconstruction effect of our SR algorithm with BTV algorithm and Tikhonov method.

Firstly, we create a sequence of LR images with a HR image of size 1232x752 pixels, and the reduction ratio is 2, 4 and 8 respectively to test the performance in different situation. The HR image is shifted, Gaussian blurred, downsampled and added Gaussian noise with the variance of 0.001 in to LR images. And we utilize the cubic spline sub-pixel interpolation in each iteration to increase the accuracy of interpolation. For our registration method, we divide the image into 3X3 blocks, and select the three blocks in the diagonal.

We use the mean absolute error (MAE) to represent the registration error in Fig.(3). According to Fig.3 (a), we find that the three methods have almost the same precision. The cost of time is shown in Fig.3 (b). The vertical axis is the average computational time of a pair of images, and the abscissa axis is the size of images. We find that our new Keren method can greatly reduce the computational time, which is 9 to 17 times more rapid than the other two methods for the image of size 635X368 pixels.

Then we perform the SR image reconstruction experiment with our edge enhancing algorithm. In the first experiment, we use the LR image sequence provided by Sina Farsiu [18]. We use the first 20 LR images in our experiment, and reconstruct the HR image with a resolution enhancement factor of 4. Fig.4 (a) is one of the LR images. Fig.4 (b) shows the cubic spline interpolation result. Fig.4 (c) shows the Laplacian

regularization result, with the regularization factor $\mu=0.5$. The resolution obviously increased but the edge is blurred to some extent. Fig.4 (d) is the SR result of BTV regularization method. Fig.4 (e) shows our edge enhancing SR reconstruction result, with $\mu=1$ and $\alpha=0.007$. By zooming in the image to see the details, we find that our method performs better than the other methods.

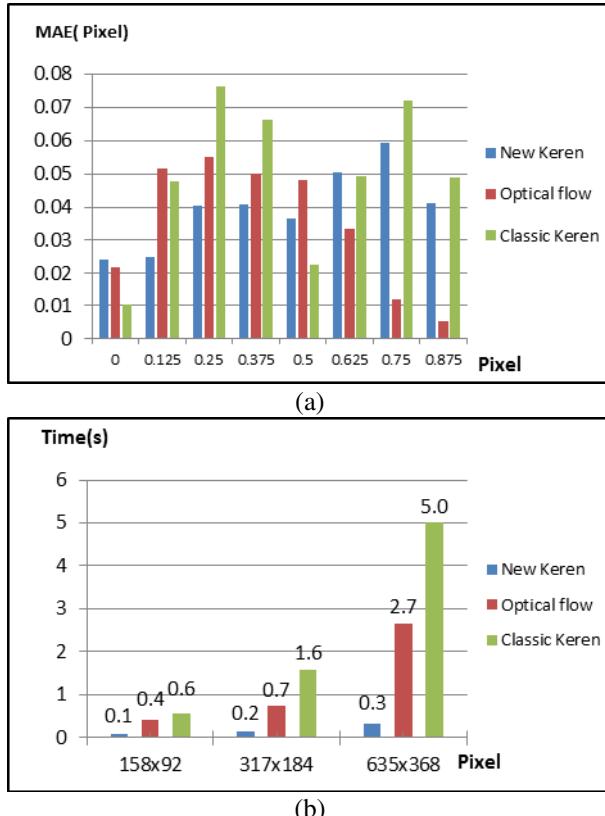


Fig. 3. The performance comparison of the three registration methods. (a) Comparison of precision. (b) Comparison of time.

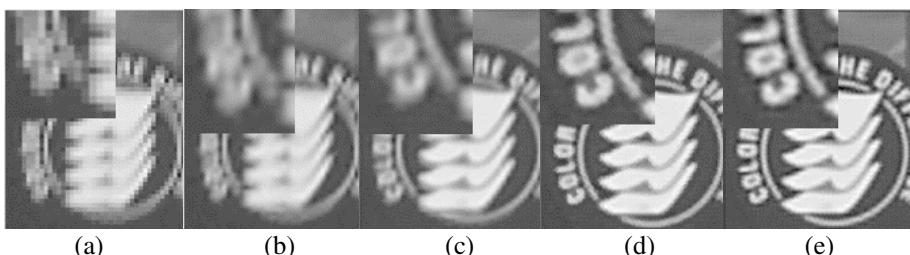


Fig. 4. Results of different SR methods for image “disk”. (a) One LR image. (b) Cubic spline interpolation. (c) Laplacian regularization. (d) BTV. (e) Edge enhancing method.

The second experiment is based on a sequence of 23 real images captured by Nikon D90 camera. We extract a sequence of 215X161 pixels parts from the entire image of the same scene. We compared the different SR reconstruction results with a resolution enhancement factor of 4. The results shown in Fig.5 are segments of the whole SR results. Fig.5 (a) is one of the LR image. Fig.5 (b) is the cubic spline interpolation results of Fig.5 (a). Fig.5 (c) uses the Laplacian regularization method, where μ is chosen to be 0.5. Fig.5 (d) is the SR result of BTV method. Fig.5 (e) is the result of our SR approach, with $\mu=3$ and $\alpha=0.01$.

From the above two experiments, we found that: our proposed method can produce better results than Laplacian regularization method, since the edge is sharper in our method. It maintains more edge information than BTV method as well as removing the noise, BTV considers the noisy pixel nearby the edge as edge information, and thus the edge is wider.

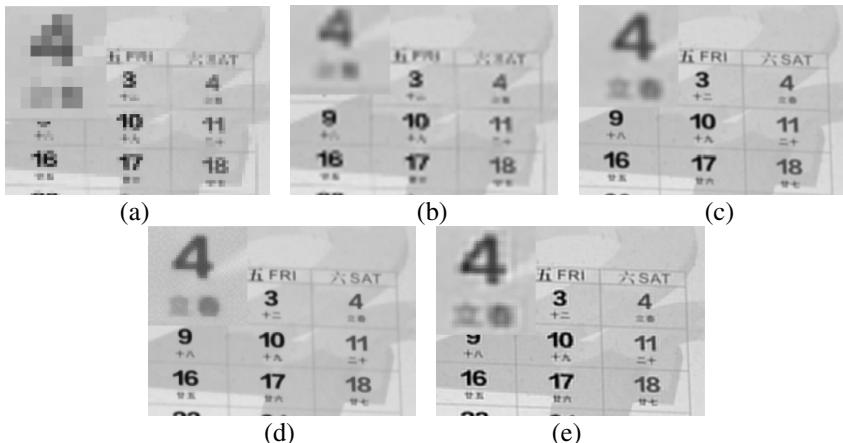


Fig. 5. Results of different SR reconstruction methods. (a) One of LR images. (b) Cubic spline interpolation. (c) Laplacian regularization. (d) BTV. (e) Edge enhancing method.

5 Conclusion

In this paper, we introduced a fast and high precise image registration method and an edge enhancing SR image reconstruction method. The image registration method is based on the Keren method. The proposed method divides the two images into several equal sub-images, and then uses a part of them to estimate the motion parameters. It has been proved the superiority in computational cost and estimation precision of our Keren based image registration method. The edge enhancing SR method is a new regularization based SR method, in which it combines two regularization components with different effects. The results indicate that the proposed method performs better in removing the noise and enhancing the edge information.

Acknowledgment. This work was partly supported by the National Natural Science Foundation of China (No. 60875072, 61273350) and Beijing Natural Science Foundation (Grant 4112035). We also give our thanks to Sina Farsiu from Duke University for providing data of images in our experiments.

References

1. Keller, J.B.: Inverse Problems. *Am. Math. Mon.* 83, 107–118 (1976)
2. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. *IEEE Signal Process. Mag.* 20, 21–36 (2003)
3. Keren, D., Peleg, S., Brada, R.: Image sequence enhancement using sub-pixel displacements. In: 1988 Proc. of the Comput. Soc. Conf. Comput. Vis. Pattern Recognit., CVPR 1988, pp. 742–746 (1988)
4. Horn, B.K.P., Schunck, B.G.: Determining optical flow (distribution of apparent movement velocities of image brightness patterns). In: Tech. Appl. Image Underst. Proc. Meet., pp. 319–331 (1981)
5. Li, R., Zeng, B., Liou, M.-L.: A new three-step search algorithm for block motion estimation. *IEEE Trans. Circuits Syst. Video Technol.* 4, 438–442 (1994)
6. Elad, M., Hel-Or, Y.: A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Trans. Image Process.* 10, 1187–1193 (2004)
7. Elad, M., Feuer, A.: Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans. Image Process.* 6, 1646–1658 (1997)
8. Huang, T., Tsai, R.: Multi-frame image restoration and registration. *Adv. Comput. Vis. Image Process.* 1, 317–339 (1984)
9. Teodosio, L., Bender, W.: Salient video stills: Content and context preserved. In: Proc. First Acm Int. Conf. Multimed., vol. 10, pp. 39–46 (1993)
10. Stark, H., Oskoui, P.: High-resolution image recovery from image-plane arrays, using convex projections. *J. Opt. Soc. Am.* 6, 1715–1726 (1989)
11. Tom, B.C., Katsaggelos, A.K.: Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. In: 1995 Proc. of the Int. Conf. Image Process., vol. 2, pp. 539–542 (1995)
12. Schultz, R.R., Stevenson, R.L.: Extraction of high-resolution frames from video sequences. *IEEE Trans. Image Process.* 5, 996–1011 (1996)
13. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. *IEEE Trans. Image Process.* 13, 1327–1344 (2004)
14. Yang, M.-C., Wang, C.-H., Hu, T.-Y., Wang, Y.-C.F.: Learning context-aware sparse representation for single image super-resolution. In: 2011 18th IEEE Int. Conf. Image Process. ICIP, pp. 1349–1352 (2011)
15. Gajjar, P.P., Joshi, M.V.: New Learning Based Super-Resolution: Use of DWT and IGMRF Prior. *IEEE Trans. Image Process.* 19, 1201–1213 (2010)
16. Tanaka, M., Okutomi, M.: A fast MAP-based super-resolution algorithm for general motion. In: Bouman, C.A., Miller, E.L., Pollak, I. (eds.), vol. 6065, pp. 404–415 (2006)
17. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Phys. Nonlinear Phenom.* 60, 259–268 (1992)
18. Farsiu, S.D.: Mdsp Super-Resolut. Demosaicing Datasets at,
<http://users.soe.ucsc.edu/~milanfar/software/sr-datasets.html>

Foreground Detection via Motion Field Based MRF-MAP

Limin Zhu and Yue Zhou

Institute of Image Processing and Pattern Recognition
Shanghai Jiaotong University
`{zlm_laker, yue_zhou}@sjtu.edu.cn`

Abstract. Foreground detection has always been of great concern in image processing field. This paper focused on detecting moving objects in videos taken by moving cameras. The segmentation result is produced by a MRF-MAP labeling method based on the motion field of optical flows. The main idea is based on the difference between foreground and background movement. Our method is evaluated on different videos to show its effectiveness.

Keywords: MRF-MAP, motion field, background, optical flows.

1 Introduction

There have been numerous studies on background subtraction methods used in surveillance system of stationary cameras. However, most of them take the assumption that the camera is stable. This is a problem becoming increasingly significant since moving camera platforms, like mobile phones and robots, are becoming part of our lives. Under this circumstance, there is a pressing need to build a system to detect foreground objects which can be used in moving cameras.

As is stated in [1], a good background subtraction technique should be able to handle the following questions: illumination changes, high frequency background objects and background movement, all of which will occur continuously in sequences taken from moving cameras.

There have been lots of background techniques available nowadays, some of which produce good effects in static cameras. The basic idea was to take use of the differences between frames which dated back to the late 70s[2]. Subsequent approaches were proposed like GMM[3], Kalman Filters[4], non-parametric kernel density estimates[5], local binary patterns[6], sample-based techniques[7,8,9,10].

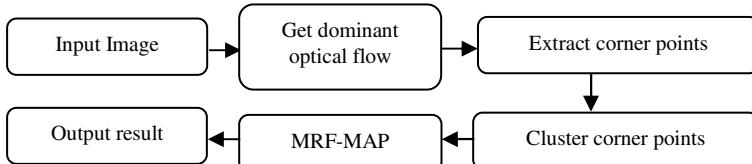


Fig. 1. The flow chart of our system

The underlying assumption of the above methods was that there existed no movement or little movement in the background. The assumption of static background led to a requisite for camera to stay stationary when the sequence is taken.

As for mobile cameras, a lot of effort has been made to relax the assumption of static background model, which has largely relied on ego-motion compensation[11,12]. Compensation is made for movement of background through a homograph or 2D affine transformation. However camera center should not translate or rotate under such circumstances which limit the camera motion to panning, tilt, or zooming. Some other methods have been developed in the case of camera center deviation. [13,14]

Another way to detect foreground pixels in a moving camera is a layer-based method which models the scene as several piece-wise planar scenes, and cluster segments based on motion coherency.

Finally, trajectories are also used often to distinguish foreground and background movement. [15,16] sparsely segment point trajectories based on geometric constraints.

In contrast to all the approaches, the method we proposed in this paper focuses on the optical flow difference between foreground and background pixels. Thus camera calibration is not needed and there is no limitation of camera motion required. A MRF-MAP based region growth method is also applied in our method.

Our paper is organized as follows: Section 2 explains our method and shows the result of clustering optical flows, Section 3 presents the foreground detection result and analysis, Section 4 concludes our paper.

2 Our Approach

In this section, we will present a background subtraction technique used for moving cameras. It is based on optical-flow clustering with reduced calculation. Though optical flow has been studied a lot, it has rarely occurred that optical flow is used to detect foreground in moving cameras because of its large computational cost and complexity of its classification. Our approach is based on the difference between optical flows of background and foreground. The movement of the camera produces a dominant optical flow which means that most of the background pixels in the current frame possess similar optical flows while those optical flows of the foreground are either different in their magnitudes or directions. The section below will explain in detail how we use optical flow to achieve foreground detection when there even exists large camera movement.

2.1 Dominant Optical Flow Extraction

The first step is to catch the dominant direction and magnitude of the optical flow caused by the camera movement. Since the calculation of the whole frame's optical-flow is too slow and difficult to cluster, we need a way to reduce the calculation burden. In this sense a set of points is chosen evenly throughout the image. Suppose

that we have got an image with the height of h and width of w , the set will consist of $h/hspace * w/wspace$ elements where $hspace$ and $wspace$ indicate the space between two nearby chosen pixels. We get larger set with smaller $hspace$ and $wspace$. This set will inevitably contain some of the foreground pixels. Lucas-Kanade is then used to calculate their optical flows.

The optical-flow of pixel x is denoted as a three element set $O(x) = \{d, s, p\}$ (d represents the direction of the optical flow, s represents the magnitude of the optical flow vector, p is the pixel's coordinate in a frame), where d is an integer between 0-9. d is 0 if the magnitude of a optical flow is too small and the rest of the bins represent the angles evenly distributed from 0 to 360 degree, which means 1 standing for 0-45 degree and 8 standing for 315-360 degree. The d th bin($k=0,1,2,\dots,8$) consists of several clusters and each of them can be denoted as $C_k(j) = \{mean_{d_j}, number_{d_j}\}$, where $mean_{d_j}$ records the mean value of the magnitude of all the optical flows in the j -th cluster, and $number_{d_j}$ records the number of elements in the j -th cluster. The optical flows are then clustered by K-means in those bins. After clustering, the dominant optical flows can be extracted.

The following algorithm shows how it works.

Algorithm 1

Input: T pixel points with its position in the frame and its optical-flow, $O(x_i) = \{d_i, s_i, p_i | i = 1, 2, 3 \dots T\}$

Output: Dominant optical flow with its magnitude and direction

Method:

1. **for** $i=1 \rightarrow T$ **do**
 - if** there are L clusters existing in the d_i th bin **do**
 - $Dist = min_j |s_i - mean_{d_{ij}}| (j = 1, 2, 3 \dots T)$
 - if** $Dist < threshold$, add $O(x_i)$ to the j -th cluster and update its $mean_{d_{ij}}$ and $number_{d_{ij}}$
 - else** create a new cluster in the d_i th bin with $O(x_i)$
 - end for**
 2. Let d_{cur} denote the bin with the most elements and $mean_{d_{cur}}$ denote the mean value of the cluster with the biggest $number_{d_{cur}}$ in the d_{cur} -th bin.
-

Then d_{cur} and $mean_{d_{cur}}$ (also denoted as μ_0 in MRF-MAP) can describe the features of the dominant optical flows in the current frame. Also the variance of the optical flows in the selected cluster is calculated and denoted as $variance_{d_{cur}}$ (also as σ_0^2 in MRF-MAP, section 2.3).

2.2 Extraction and Clustering of Corner Points

Foreground objects possess optical flows with different directions or magnitude compared with the dominant optical flow caused by camera movement. Theoretically

speaking, it is feasible to calculate all the optical flows of the image and label the pixels accordingly. However the fact is that it is both time consuming and inaccurate due to the noise existing in the frame. Since moving objects have clear boundaries for the most of the time, our foreground detection is therefore based on the clustering of corner points first. Harris corner detector is used to get those corner points, after which is a clustering procedure to eliminate those belonging to background. The algorithm below shows the clustering stage.

Algorithm 2

Input: T corner points with their coordinates in the frame and optical-flows, $O(x_i) = \{d_i, s_i, p_i | i = 1, 2, 3 \dots T\}$

Output: Corner points with special optical-flows. $O(x_l) = \{d_l, s_l, p_l | 0 < l < T\}$

Method:

```

1. for i=1→T do
    if  $d_i \neq d_{cur} \text{ || } |s_i - mean_{cur}| > threshold$  do
        add  $O(x_i)$  to  $O(x_l)$ 
    end for

```

Fig.2 shows the result after those corner points are selected and clustered.



Fig. 2. From left to right are: a) input image, b) optical flow of feature points before clustering, c) optical flows chosen after clustering.

Since $O(x_l)$ only consists of some of the foreground pixels, it is necessary to apply a certain kind of region growth method to it. However it is not necessary to apply it to every single corner point since the surrounding regions of corner points might be overlapping. Therefore nearby corner points are therefore clustered into a group for later use. Here is what we are going to do: we cluster $O(x_l)$ according to their p_l , each set after clustering is denoted as $L_m = \{center_m, radius_m, n_m\}$ ($center_m$ is the center of the circle, $radius_m$ is the radius of the circle, n_m is the total number of elements in the m-th cluster). The clustering process is based on the Euclidian distance between $center_m$ and the current pixel. The $radius_m$ is updated as the longest distance from the point in the cluster to its $center_m$.

After the clustering procedure, we can get some circles covering those pixels close to each other.

The following algorithm shows how it works:

Algorithm 3

Input: K corner points with their $O(x_l) = \{d_l, s_l, p_l | l = 1, 2, 3 \dots, K\}$.

Output: N circles covering corner points

Method:

1. initialize $L_1 = \{p_1, 0, 1\}$

2. **for** $l=2 \rightarrow K$ **do**

 Suppose we have M existing clusters

$\text{Dist} = \min_m \|p_l - \text{center}_m\|_2^2, (m = 1, 2, 3 \dots M)$

if $\text{Dist} < \text{threshold}$ **do**

 add p_l to the m-th cluster and update its parameters.

$\text{center}_m = (\text{center}_m \times n_m + p_l) / (n_m + 1)$

$R = \|p_l - \text{center}_m\|_2^2$,

if $R > \text{radius}_m$, $\text{radius}_m = R$

n_m++

else create a new cluster with p_l

end for

2.3 MRF-MAP Labeling

The approach introduced in this section is aimed to detect foreground object via searching from clustered corner point sets after section 2.2. A nearby pixel should be treated as foreground if its optical flow has a unique direction or magnitude. Once a pixel is labeled as foreground due to its direction of the optical flow, it will not change until next frame. MRF-MAP is used to label those pixels with unique magnitude of their optical flows.

MRF-MAP is widely used in visual labeling which in our case is a 0 or 1 question, where 0 represents the pixel belonging to background and 1 the foreground.

The unlabeled region is denoted as $X = \{x_1, x_2, x_3, \dots, x_N\}$, in which x_i stands for the i-th pixel in the region and N is the total number of pixels in this area. $Y = \{y_1, y_2, y_3, \dots, y_N\}$ represents the labels of pixels in which y_i is the label of pixel x_i and $y_i \in \{0, 1\}$. The labeling question is then turned into a MAP problem (maximum a posterior):

$$P(X|Y) = \frac{1}{Z(T)} \cdot e^{-E(X)/T} \quad (1)$$

where T is a constant called the temperature which shall be assumed to be 1 unless otherwise stated, and $E(X)$ is the energy function. To maximize $P(X|Y)$ is equal to minimizing $E(X)$ which is defined as:

$$Y^* = \operatorname{argmin} E(X) \quad (2)$$

$$E(X) = \sum_{i \in C_1} E_1(x_i | \theta_k, y_i) + \sum_{\{i, i'\} \in C_2} E_2(y_i, y_{i'}) \quad (3)$$

$$E_1(x_i | \theta_k, y_i) = (s_{x_i} - \mu_{y_i}) / 2 \sigma_{y_i}^2 \quad (4)$$

$$E_2(y_i, y_{i'}) = \begin{cases} 1, & y_i = y_{i'} \\ 0, & y_i \neq y_{i'} \end{cases} \quad (5)$$

$$\theta_k = \{\mu_0, \sigma_0^2; \mu_1, \sigma_1^2\} \quad (6)$$

C_1 is a single site and C_2 is a pair of 4 neighboring sites. s_{x_i} is the magnitude of the optical flow of pixel x_i .

We have already got μ_0, σ_0^2 in section 2.1 and N clusters after 2.2. For each cluster, all the optical flows of pixels within $radius_m$ from its $center_m$ are computed and labeled as foreground if its $d_i! = d_{cur} \parallel |s_i - mean_{cur}| > threshold$, afterwards the mean and variance of all foreground pixels are calculated and denoted as μ_1, σ_1^2 . One advantage for this approach is that the accuracy of ICM solution for solving MRF-MAP mainly depends on the initialized labels of the elements and by doing this we assign a more plausible label to each pixel at the beginning.

Then ICM is used to solve MAP-MRF in our approach, the iteration ends if the change of the $E(X)$ is less than 1% between two iterations.

3 Experimental Results

Generally speaking, the main difference between background extraction for moving camera and static camera is that the movement of background objects existing in mobile cameras will add lots of trouble to the foreground detection. But the overall optical flow of background is similar and thus making it possible for us to distinguish those optical flows of those actual moving foreground objects since they tend to move in a different way either in direction and speed. With MRF-MAP, a more accurate labeling result can be produced.

Our method is evaluated on several sequences taken by hand-held cameras. The first one is taken from a view of two-storey-high. There is high-frequency change like light and waving plants in the video. The second one is taken from a traffic crossroad. Fig.2 explicitly shows our results.

The result is self-explanatory which shows that despite large camera movement and illumination change in the video, our method is still able to tell foreground pixels from background movement. The reason for that is optical flow produced by foreground pixels are either unique in its direction or magnitude.



Fig. 3. The first and third rows are the input videos. The second and fourth are the corresponding results.



Fig. 4. The first row is the input video. The second is the result output by [17] the third row is the result produced by our method.

4 Conclusion

This paper has proposed a novel method of background subtraction in mobile cameras. Unlike other approaches studied before, it mainly concentrates on the optical

flow clustering. Optical flow is seldom used in background subtraction due to its computation complexity, but its weakness is overcome by selecting a set of feature points. A MRF-MAP based region growth method is introduced to get all of the foreground pixels and make the result more satisfactory. Our method is evaluated on different sequences and proven to be effective.

References

- [1] Piccardi, M.: Background subtraction techniques: A review. In: Proc. IEEE Int. Conf. Syst., Man Cybern., The Hague, The Netherlands, vol. 4, pp. 3099–3104 (2004)
- [2] Jain, R., Nagel, H.: On the analysis of accumulative difference pictures from image sequences of real world scenes. IEEE TPAMI (1979)
- [3] Stauffer, C., Grimson, E.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 747–757 (2000)
- [4] Koller, D., Weber, J., Huang, T., Malik, J., Ogasawara, G., Rao, B., Russell, S.: Towards robust automatic traffic scene analysis in real-time. In: ICPR (1994)
- [5] Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE (2002)
- [6] Heikkila, M., Pietikainen, M.: A Texture-Based Method for Modeling the Background and Detecting Moving Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4) (April 2006)
- [7] Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
- [8] Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters 27(7), 773–780 (2006)
- [9] Wang, H., Suter, D.: A consensus-based method for tracking: Modelling background scenario and foreground appearance. Pattern Recognition 40(3), 1091–1105 (2007)
- [10] Barnich, O., Van Droogenbroeck, M.: ViBe: A Universal Background Subtraction Algorithm for Video Sequences. IEEE Transactions on Image Processing 20(6) (June 2011)
- [11] Ren, Y., Chua, C.-S., Ho, Y.-K.: Statistical background modeling for non-stationary camera. Pattern Recogn. Lett. 24(1-3) (2003)
- [12] Hayman, E., Eklundh, J.-O.: Statistical background subtraction for a mobile observer. In: IEEE ICCV (2003)
- [13] Tao, H., Sawhney, H.S., Kumar, R.: Object tracking with Bayesian estimation of dynamic layer representations. IEEE TPAMI (2002)
- [14] Xiao, J., Shah, M.: Motion layer in the presence of occlusion using graph cuts. IEEE TPAMI (2005)
- [15] Kanatani, K.: Motion segmentation by subspace separation and model selection. IEEE ICCV (2001)
- [16] Vidal, R., Ma, Y.: A unified algebraic approach to 2-D and 3-D motion segmentation. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 1–15. Springer, Heidelberg (2004)
- [17] Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV 2009 (2009)

The Speaker Recognition of Noisy Short Utterance

Ying Chen and Zhen-Min Tang

School of Computer Science & Engineering,
Nanjing University of Science and Technology, Jiangsu Nanjing 210094

Abstract. The noisy short utterance is polluted by noise and its corpus is not full, so the recognition rate significantly decreased. This paper proposed noise separation algorithm based on constrained Non-negative matrix factorization (CNMF), use it to separate pure speech from noisy speech. And then the speech frames are classified to high quality class and low quality class using differences detection and discrimination algorithm (DDADA) proposed in this paper. Combining features group with GMM-UBM two-stage classification model to make full use of limited information. Experiments show that the above algorithms improve speaker recognition rate of noisy short utterance.

Keywords: speaker recognition, noisy short utterance, CNMF, DDADA, features group, GMM-UBM two-stage classification model.

1 Introduction

The field of speaker recognition has made a lot of research achievements [1], the rate of speaker recognition under laboratory environment (clean speech) is high. Mel frequency cepstral coefficients and universal background model UBM is widely used, document [2] extends the application of UBM model in noisy environment. Joint factor analysis technology recently proposed by Kenny (JFA) has opened up a new direction for the study of speaker recognition under channel mismatch, on the basis of this, Dehak proposed the concept of I- vector (Identity Vector, I-Vector) [3].

The research results show a linear predictive residual signal contains useful information of speech signal[4-5], document [6] confirmed WOCOR compensates for MFCC. Document [7] confirmed the identification results of WOCOR is better when speech content in training and testing is mismatch, and the combination of WOCOR and MFCC obtain better recognition effect, but the recognition rate were significantly lower in noise environment. So we proposed the following algorithm to improve speaker recognition rate: CNMF, DDADA, combining features group with GMM-UBM two-stage classification model.

2 CNMF and DDADA

Recently, many researchers add constrains in NMF and use it to separate noise and speaker recognition [8-9]. We use FastICA to initialize NMF, and add discriminating constrain to NMF.

Given a non-negative matrix $W = [x_1, x_2, \dots, x_n] \in R^{m \times n}$, W can be expressed as the product of two non-negative matrices, $W=UV$.

Assuming that the first l samples are labeled in the data set $\{x_i\}_{i=1}^n$, the remaining $n-l$ samples are not labeled, the data set consists of c samples class. If the sample x_i belongs to the j th class sample, then $c_{ij} = 1$, otherwise $c_{ij} = 0$, so we define a constrained classification matrix $A = \begin{pmatrix} C_{l \times c} & 0 \\ 0 & I_{n-l} \end{pmatrix}$, I_{n-l} is a $(n-l)(n-l)$ -dimensional unit matrix. Add classification information to non-negative matrix Z , we can get $V = AZ$, where $Z \in R^{c+n-l \times k}$, $k \ll m$ and $k \ll n$. If x_i and x_j belong to the same sample class, then $v_i = v_j$, and $X \approx U(AZ)^T$.

We can get speech frame energy spectrum of the input signal through the FFT. Therefore, we use a Multi-dimension Gauss distribution $S(\boldsymbol{\alpha}, \Sigma)$ to describe the spectral character of speech signal, where, $\boldsymbol{\alpha}$ is the mean vector of speech frame energy, Σ is the covariance matrix. In order to reduce the computation cost, assumed Σ to be diagonal matrix, the speech model can be expressed as $S(\boldsymbol{\alpha}, \sigma^2)$.

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_N)' \quad (1)$$

$$\sigma^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \dots, \sigma_N^2)' \quad (2)$$

Before testing, reserve several frames as pure speech to initialize the detection model. After that, calculate the similarity evaluation of each speech frame. If the spectrum features of input frame is similar to pure speech, then the similarity evaluation of the frame is higher, otherwise, is lower. Evaluation of each frame can be expressed as:

$$score(O_i) = S(O_i; \boldsymbol{\alpha}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left| -\frac{(O_i - \boldsymbol{\alpha})^2}{\sigma^2} \right| \quad (3)$$

In actual calculation, use the formula (4) to instead of (3).

$$score(O_i) = \frac{(O_i - \boldsymbol{\alpha})^2}{\sigma^2} + \ln \sigma^2 = \sum_{n=1}^N \left| \frac{(O_{i,n} - \alpha_n)^2}{\sigma_n^2} + \ln \sigma_n^2 \right| \quad (4)$$

$O_i = (O_{i,1}, O_{i,2}, O_{i,3}, \dots, O_{i,J})'$ is the energy spectrum vector of the current frame. If the similarity evaluation of current frame is low, then the frame is classified to low quality class, otherwise the current frame is classified to high quality class. High quality frames is used to update the detection model. The update process is an iterative process, which makes the detection model be more close to the pure speech model. The update process can be expressed as:

$$\boldsymbol{\alpha}_{m+1} = \frac{m\boldsymbol{\alpha}_m + S_{m+1}}{m+1} \quad (5)$$

$$\sigma_{m+1}^2 = \frac{(m-1)\sigma_m^2 + (S_{m+1} - \bar{\sigma}_m)^2}{m} - (\bar{\sigma}_{m+1} - \bar{\sigma}_m)^2 \quad (6)$$

$\bar{\sigma}_m$ 、 σ_m^2 are respectively mean vector and variance vector before updated; $\bar{\sigma}_{m+1}$ 、 σ_{m+1}^2 are respectively mean vector and variance vector after updated; m is the speech frames before updated; S_{m+1} is the energy spectral vector of speech frame after updated.

3 Extracting Robust Features

In this paper, we combine WOCOR with MFCC_D_LPCC. WOCOR Feature Parameters have K octave groups of wavelet coefficients,

$$\mathbf{W}_k = \left\{ \mathbf{w}(2^k, b) \mid b = 1, 2, \dots, N \right\}, k = 1, 2, \dots, K \quad (7)$$

Each octave group of coefficients is divided evenly into M subgroups

$$\mathbf{W}_k^M(\mathbf{m}) = \left\{ \mathbf{w}(2^k, b) \mid b \in \left(\frac{(m-1)N}{M}, \frac{mN}{M} \right] \right\}, m = 1, 2, \dots, M \quad (8)$$

The complete feature vector is composed of $K \cdot M$ parameters, as follows:

$$\mathbf{WOCOR}_M = \left\{ \|\mathbf{W}_k^M(\mathbf{m})\| \mid k = 1, 2, \dots, K, m = 1, 2, \dots, M \right\} \quad (9)$$

In this paper, we let $K = 6$ to deal with speech data in the frequency range of 0 to 4000 H, $M=4$, $M=6$, $M=8$, respectively.

MFCC (Mel Frequency Cepstrum Coefficient) and LPCC(Linear Prediction Cepstrum Coefficient) are the commonly parameters in the speaker recognition. We added Delta features to the above two kinds of features, composed MFCC_D_LPCC feature. We get three subsystems: MFCC_D_LPCC+WOWOR4, MFCC_D_LPCC+WOWOR6, MFCC_D_LPCC+WOWOR8, the highest score of three subsystems is the final score of the system.

4 GMM-UBM Two-Stage Classification Model

GMM is used to estimate the D-dimensional feature vector \vec{x} for the task of speaker recognition. Assuming K diagonal Gaussian mixture components, the probability density function of a GMM is given by (10):

$$p(\vec{x} \mid \lambda) = \sum_{c=1}^K w_c \prod_{m=1}^D N(x_m, \bar{\sigma}_{c,m}, \sigma_{c,m}^2) \quad (10)$$

Where w_c is the component weight and $N(x_m, \infty_{c,m}, \sigma_{c,m}^2)$ is a uni-variate Gaussian distribution with mean $\infty_{c,m}$ and variance $\sigma_{c,m}^2$.

$$N(x_m, \infty_{c,m}, \sigma_{c,m}^2) = \frac{1}{\sqrt{2\pi\sigma_{c,m}^2}} \exp\left(-\frac{(x_m - \infty_{c,m})^2}{2\sigma_{c,m}^2}\right) \quad (11)$$

$$\lambda = (w_c, \vec{\infty}_c, \vec{\sigma}_c^2) \quad c = 1, \dots, K. \quad (12)$$

Feature vectors \vec{x} is classified into two sub vectors, which are reliable R and unreliable U . In the process of recognition, the two sets are used for speaker recognition respectively. Reliable R is used directly to estimate the similarity score of the speaker λ . We assume that unreliable components are polluted by additive noise, but they do contain information about the maximum energy of the target speech component, so deal with them for recognition.

$$p(\vec{x}|\lambda) = \sum_{c=1}^K w_c \prod_{r \in R} N(x_r, \infty_{c,r}, \sigma_{c,r}^2) \cdot \prod_{u \in U} \frac{1}{x_{high,u} - x_{low,u}} \int_{x_{low,u}}^{x_{high,u}} N(x_u, \infty_{c,u}, \sigma_{c,u}^2) dx_u \quad (13)$$

The bounds were set to $[x_{low,u} - x_{high,u}] = [0, x_u]$. The integral in (13) can be evaluated as the vector difference of error function, and (13) can be rewritten as (14).

$$p(\vec{x}|\lambda) = \sum_{c=1}^K w_c \prod_{r \in R} N(x_r, \infty_{c,r}, \sigma_{c,r}^2) \cdot \prod_{u \in U} \frac{1}{x_{high,u} - x_{low,u}} \frac{1}{2} \left[\operatorname{erf}\left(\frac{x_{high,u} - \infty_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right) - \operatorname{erf}\left(\frac{x_{low,u} - \infty_{c,u}}{\sqrt{2\sigma_{c,u}^2}}\right) \right] \quad (14)$$

5 Experiments and Results

5.1 Speech Database and Noise

The speech database is the TIMIT speech database, the sampling rate is 16KHz, mono recording, 16Bit quantification, including 630 speaker, contains two subdirectories: train directory and test directory. Each directory contains 8 folders, the eight folders represent eight different dialects of English, each speaker read 10 statements, and the length of each sentence is about 3 seconds. The complex noise under battlefield environment was added to each speech according to different SNR. Experimental samples were obtained from TIMIT speech database added noise, and we used 230 speakers of them; the training corpus was taken from the first sentence of each speaker, the testing corpus was taken from each speaker's tenth words.

5.2 Research on Noise Separation Algorithm

We used CNMF to separate pure speech from noisy speech, and then used three methods to make the following experiment. Method 1: combine pure speech with

GMM-UBM model. Method 2: use DDADA to get high quality speech frame and low quality speech frame, delete low quality speech frame, combine high quality speech frame with GMM-UBM model. Method 3: use DDADA to get high quality speech frame and low quality speech frame, use high quality speech frame to speaker recognition directly, use low quality speech frame processed to speaker recognition, combine them with GMM-UBM two-stage classification model. In the feature extraction phase, we used MFCC、MFCC_D_LPCC、WOWOR4 respectively.

Table 1. Combine MFCC with the above three methods respectively

methods number	SNR (dB)			
	-5	0	5	10
method 1	36.521%	43.913%	49.130%	50.869%
method 2	41.739%	49.130%	54.782%	56.086%
method 3	47.391%	52.608%	56.956%	58.260%

Table 2. Combine MFCC_D_LPCC with the above three methods respectively

methods number	SNR (dB)			
	-5	0	5	10
method 1	38.695%	45.652%	51.304%	53.043%
Method 2	45.652%	51.739%	56.956%	58.695%
method 3	52.173%	57.826%	61.304%	62.608%

Table 3. Combine WOWOR4 with the above three methods respectively

methods number	SNR (dB)			
	-5	0	5	10
Method 1	31.304%	37.826%	41.739%	45.217%
method 2	37.391%	43.043%	47.391%	50.434%
method 3	41.304%	46.956%	51.304%	53.043%

5.3 The Relationship between Recognition Rate and Feature Group

Table 4. Combine feature groups with the method 3

feature groups	SNR (dB)			
	-5	0	5	10
MFCC + WOWOR4	53.043%	58.260%	60.434%	61.739%
MFCC_D_LPCC+ WOWOR4	56.956%	60.869%	63.043%	64.347%
(MFCC_D_LPCC+WOWOR4)+				
(MFCC_D_LPCC+WOWOR6)+	61.739%	64.782%	66.521%	67.391%
(MFCC_D_LPCC+ WOWOR8)				

6 Conclusions

In this paper, we used CNMF to separate more pure corpus from noisy short utterance, and used DDADA to classify the speech frames into high quality class and low quality class accurately, finally combined features group with GMM-UBM two-stage classification model. The experiments confirmed that the above algorithms can improve speaker recognition rate of noisy short utterance.

References

1. Fatima, N., Zheng, T.F.: Short utterance speaker recognition. In: International Conference on Sustems and Informatics(ICSAI), pp. 1746–1750 (2012)
2. May, T., van de Par, S., Kohlrausch, A.: Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Transactions on Audio, Speech and Language Processing 20, 108–121 (2012)
3. Dehak, N., et al.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio,Speech and Language Processing 19, 788–798 (2011)
4. Murty, K.S., Yegnanarayana, B.: Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process 13, 52–55 (2006)
5. Chetouani, M., Faundez-Zanuy, M., Gas, B., Zarader, J.L.: Investigation on LP-residual representations for speaker identification. Pattern Recognition 42, 487–494 (2009)
6. Zheng, N., Ching, P.C., Lee, T.: Time frequency analysis of vocal source, signal for speaker recognition. In: Proc. ICSLP, pp. 2333–2336 (2004)
7. Chen, W.N., Zheng, N., Lee, T.: Discrimination power of vocal source and vocal tract related features for speaker segmentation. IEEE Transactions on Audio, Speech and Language Processing 15, 1884–1892 (2007)
8. Joder, C., Weninger, F., Eyben, F., Virette, D., Schuller, B.: Real-time speech separation by semi-supervised nonnegative matrix factorization. In: Proc. of Inter. Conf Latent Variable Analysis and Signal Separation (2012)
9. Joder, C., Schuller, B.: Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition. ITG-Fachbericht 236: Sprachkommunikation, Braunschweig (2012)

Hyper-graph Matching with Bundled Feature

Deyuan Li and Yue Zhou

Institute of Image Processing and Pattern Recognition
Shanghai Jiao Tong University
Shanghai, China
deyuanli2010@gmail.com, zhouyue@sjtu.edu.cn

Abstract. Hyper-graph matching algorithm describes the whole structure of object by high-order topology. Previous work has presented many methods to build and solve the problem model. This paper mainly focuses on feature description and result optimization. First, we combine several features in stable region of object as bundled feature, it can describe more relationship by hyper-graph model. Second, we properly extend previous work to build and solve the hyper-graph model. Finally, we optimize the matching result by iteration and modifying constraints, it improves the accuracy effectively. Comparative experiments verify the good performance of our algorithm, especially for non-rigid object matching.

Keywords: Hyper-graph matching, non-rigid object matching, bundled feature, result optimization.

1 Introduction

Object matching is a basic problem in pattern recognition. It has extensive uses in object detection and tracking [1], image classification [2], image retrieval [3] and so on. Object matching problem is mainly related to two aspects: (i) feature extraction and description [4–6]. (ii) similarity calculation [1, 7–9]. Hyper-graph matching algorithm describes the whole structure of object by high-order topology. The association among points (or regions) can reduce the mismatch rate effectively. Previous work has presented many hyper-graph methods [10–12]. Because hyper-graph matching problem is NP-hard, many researches study on modeling and relaxation. R. Zass and A. Shashua [10] derived the hyper-graph matching problem in a probabilistic setting represented by a convex optimization. O. Duchenne et al. [11] defined the hyper-model by a tensor representing the affinity between feature tuples. Jungmin Lee et al. [12] reinterpreted the random walk concept on the hyper-graph in a probabilistic manner. However, most existing literatures use physical location of points to describe the relationship. Different from them, this paper uses bundled feature based on region to describe more relationship, not just physical location. Besides, some features extracted from region are not sensitive to local deformation. Therefore, our algorithm has good performance on non-rigid object matching. Because of the association, partial shelter and wrong matching

will impact similarities of others. To solve the problem, we optimize the preliminary matching results. Comparative experiments verify the good performance of our algorithm, especially for non-rigid object matching.

Many objects observed in the real world are non-rigid. Non-rigid object matching is important for computer vision. It can contribute to clothing matching and pedestrian tracking. For non-rigid object matching problem, Juho kannala et al. [13] presented a non-rigid quasi-dense matching method based on match propagation algorithm. Xiang Bai et al. [14] presented a shape-based algorithm, they used skeleton information to describe the structure of object. However, it is difficult to extract stable structure from some non-rigid objects, e.g. patterns on cloth.

1.1 Related Work

Our work is closely related to approaches in [1, 7–9, 15]. Zhong Wu et al. [15] proposed bundled feature which combines SIFT [16] points in MSER [17] region. Hongsheng Li et al. [7, 8] presented a matching algorithm based on linear programming and locally affine invariant geometric constraint. The algorithm uses mathematical programming technique to match object based on SIFT feature and geometric structure of SIFT points. In order to solve the mathematical programming problem, they used the method in [1] to convert it to a linear programming problem. Lei Xu et al. [9] replaced SIFT feature in [7, 8] with MSER feature in order to improve robustness. However, [7–9, 15] only considered SIFT feature or MSER feature. We extend the approach in [7, 8] to match our bundled feature based on region, and more kinds of features can be considered. The model in [7, 8] describes the geometric structure of SIFT points, actually it describes the proportional relationship of position in spatial domain. If we bundle proper features, we can describe more relationship. In addition, [7–9] get the final matching result by a simple threshold. In order to improve the accuracy, we optimize the matching result by iteration and modifying constraints.

Our main contributions are: (i) we combine several kinds of features in stable region of object as bundled feature, it can describe more relationship by hyper-graph model, not just physical location. (ii) by introducing coefficients and definitions, we properly extend the mathematical programming approach in [6, 7] to calculate the similarity of our relationship description. (iii) we optimize the matching result by iteration and modifying constraints, it improves the performance of resisting partial shelter.

In Sect. 2, our hyper-graph matching algorithm will be introduced. In Sect. 3, the performance of our algorithm will be verified by comparative experiments. Finally, we will summarize in Sect. 4.

2 Hyper-graph Matching Algorithm

Our algorithm will be introduced in three parts: (i) feature extraction and description. We combine several kinds of features in stable region of object as

bundled feature, it can describe more relationship by hyper-graph model, not just physical location. (Sect. 2.1). (ii) model and solution (Sect. 2.2). By introducing coefficients and definitions, we properly extend the approach in [1, 7, 8] to build and solve the hyper-graph model. (iii) result optimization (Sect. 2.3). We optimize the matching result by iteration and modifying constraints, it improves the accuracy effectively.

2.1 Bundled Feature

Firstly, we need segment some regions from the given image. These regions should be stable (it may still exist in scene image) and typical (it should represent the characteristic of object). In this paper, we use MSER [17] to get stable regions.

We combine some kinds of features in region and represent them as bundled feature. For example, if we want to combine position, color and area information of a region, we can get a feature vector $F_v = (X_v, Y_v, R_v, G_v, B_v, A_v)$, where (X_v, Y_v) are the coordinate of MSER center. (R_v, G_v, B_v) are each mean value of red, green and blue in the region. A_v is the number of pixels in the region. Different kinds of features can be bundled to solve different object matching problems.

Most hyper-graph matching algorithms [10–12] use triangle which is composed of three feature points as topological structure. Different from them, we use star structure proposed in [7, 8] to describe the relationship among regions. It can be described as follows:

$$R_o = \theta_1 R_1 + \theta_2 R_2 + \dots + \theta_b R_b \quad (1)$$

where R_o is bundled feature of region o . Regions $1, 2, \dots, b$ are neighbors of region o . b is the number of neighbors of region o . θ_b is proportionality coefficient. Because the order of regions matters, the hyper-graph is directed.

By star topology of our bundled feature, we can describe more relationship, not just physical location in most researches [7, 8, 10–12]. If we bundle proper features (e.g. position, color, and area), it can resist linear variance (e.g. if we magnify and move the image, the relationship of area will not change; the color information is not sensitive to luminance variation). The resistant can be easily proved by Eqn. (1).

Feature description in our algorithm contains two aspects: (i) region term: it describes the characteristic of region based on bundled feature. (ii) relation term: it describes the relationship among regions based on bundled feature and topological structure. Note that the bundled feature which describes region term can be different from the bundled feature which describes relation term. The difference between them can make our algorithm more flexible. For example, because of affine transformation, area information in region term is useless; because relation term is mainly used to resist linear variance, some texture features are needless in relation term.

2.2 Hyper-graph Model and Relaxation

In this subsection, we will extend the model in [7, 8] to match our region term and relation term. Our bundled feature has more kinds of features and higher dimensions than [7, 8]. Consequently, we define some weight coefficients (e.g. α_n and β_m), the feature matching cost ($s_n(p, q)$) and the number of neighbours. With our modification, we can get the problem model based on [7, 8]:

$$\hat{m} = \arg \min_m \sum_{i=1}^{|A|} \left\{ \sum_{n=1}^{|F_s|} \alpha_n s_n(a_i, m(a_i)) + \lambda \sum_{m=1}^{|F_r|} \beta_m r_m(a_i, N_{a_i}; m(a_i), N_{m(a_i)}) \right\} \quad (2)$$

where $\sum_{n=1}^{|F_s|} \alpha_n s_n(a_i, m(a_i))$ represents the matching cost of region term, and

$\sum_{m=1}^{|F_r|} \beta_m r_m(a_i, N_{a_i}; m(a_i), N_{m(a_i)})$ represents the matching cost of relation term.

λ controls the relative weight between region term and relation term. $|A|$ stands for the number of regions in model image. $m(\cdot)$ is the matching function, it represents the matching result between model image and scene image. $a_i = \{feature_l | l = 1, 2, \dots, f\}$ is bundled feature which describes region i in model image. $m(a_i)$ is the region in scene image which is related to region i in model image. $s_n(p, q)$ is the feature matching cost between the region p in model image and region q in scene image based on feature n . N_{a_i} stands for the set of ordered regions in the neighborhood of a_i . $r_m(\cdot)$ is the relationship matching cost that measures the similarity between two sets of ordered regions $\{a_i, N_{a_i}\}$ and $\{m(a_i), N_{m(a_i)}\}$ based on feature m . α_n is the weight of matching cost of feature n in region term. β_m is the weight of matching cost of feature m in relation term. $|F_s|$ stands for the dimension of bundled feature in region term, $|F_r|$ stands for the dimension of bundled feature in relation term.

We define the feature matching cost as follows:

$$s_n(p, q) = \| F_{p,n} - F_{q,n} \|_1 / F_{p,n} \quad (3)$$

where $F_{p,n}$ is the n th feature of region p . We can calculate C which is a $n_m \times n_s$ feature matching cost matrix by $\sum_{n=1}^{|F_s|} \alpha_n s_n(a_i, m(a_i))$ in Eqn. (2) and Eqn. (3).

$C(h, k)$ is the feature matching cost between region h in model image and region k in scene image. We define the neighborhood of region A as j regions whose center coordinates are close to A 's. Based on Eqn. (1) and [7, 8], the relationship can be represented as follows:

$$a_i = \sum_{a_j \in N_{a_i}} W_{ij} a_j \quad (4)$$

where W is a $n_m \times n_m$ weight matrix recording proportional relationship between a region and its neighborhood, and W_i is the i th row of W recording the relational combination coefficients for a_i . If $a_j \notin N_{a_i}$, $W_{ij} = 0$. We normalize W_j and introduce the constraint $\sum_j W_j = 1$. In order to get the unique solution of W ,

we need j (j is the number of neighborhood) equations of W . Consequently, except normalization constraint, the dimension of bundled feature in relation term should be $j - 1$, therefore, the number of neighbors should be $|F_s| + 1$. However, if we bundle too many features in relation term, one region will be represented by many neighbouring regions, the redundant representation will degrade the performance. After obtaining j neighbors for every region, we can calculate the weight matrix W with Eqn. (4) and normalization constraint. Based on [7, 8], we can represent $r_m(a_i, N_{a_i}; m(a_i), N_{m(a_i)})$ in Eqn. (2) as follows:

$$\arg \min_m \sum_{i=1}^{|A|} \|m(a_i) - \sum_{a_j \in N_{a_i}} W_{ij} m(a_j)\|_1 \quad (5)$$

After computing C and W (Eqn. (5) can be solved by Eqn. (7) which will be introduced later), the solution to Eqn. (2) can be solved by

$$\begin{aligned} \min f(X) &= \text{tr}(C^T X) + \lambda |(I_{n_m} - W)XSB| \\ \text{s.t. } &X \in \{0, 1\}^{n_m \times n_s} \\ &X1_{n_s} = 1_{n_m}, X^T 1_{n_m} = 1_{n_s} \end{aligned} \quad (6)$$

where X is a $n_m \times n_s$ binary assignment matrix. $X(h, k) = 1$ means region k in scene image is the matching result of region h in model image. tr is the trace of matrix, I_{n_m} is an $n_m \times n_m$ identity matrix, $|\cdot|$ is the summation of the absolute values of all elements in matrix, n_m is the number of regions in model image, n_s is the number of regions in scene image, S is the $n_s \times f$ matrix, which records the relationship of regions in scene image, f is the dimension of bundled feature in relation term. B is a $f \times f$ weight matrix, it is a diagonal matrix whose diagonal elements are $\{\beta_1, \beta_2, \dots, \beta_f\}$. The objective function $f(X)$ consists of region term $\text{tr}(C^T X)$ which is the matrix form of the first term in Eqn. (2) and relation term $(I_{n_m} - W)XSB$ which is the matrix form of the second term in Eqn. (2).

However, Eqn. (6) is NP-hard. We can use the method in [1] to simplify it:

$$\begin{aligned} \min \sum_{i=1}^N |x_i| &\Leftrightarrow \min \sum_{i=1}^N x_i^+ \\ \text{s.t. } &x_i \leq x_i^+, x_i \geq -x_i^+, x_i^+ \geq 0 \end{aligned} \quad (7)$$

where x_i^+ is the i th auxiliary variable representing the upper bound of x_i . Based on [7, 8], we relax constraints of $X \in \{0, 1\}^{n_m \times n_s}$, $X1_{n_s} = 1_{n_m}$, $X^T 1_{n_m} = 1_{n_s}$

to

$$X \in [0, 1]^{n_m \times n_s}, X^T 1_{n_m} \leq u_{n_s}, X 1_{n_s} = 1_{n_m} \quad (8)$$

Generally, u is 1. Finally we can get the linear programming form of Eqn. (6):

$$\begin{aligned} \min f(X) &= \text{tr}(C^T X) + \lambda 1_{n_m}^T X^+ 1_{|F_r|} \\ \text{s.t. } &X \geq 0, X^+ \geq 0 \\ &(I_{n_m} - W) X S B \leq X^+ \\ &(I_{n_m} - W) X S B \geq -X^+ \\ &X^T 1_{n_m} \leq u_{n_s}, X 1_{n_s} = 1_{n_m} \end{aligned} \quad (9)$$

where $|F_r|$ is the dimension of bundled feature in relation term, X^+ is $n_m \times |F_r|$ auxiliary variable matrix. After we find out the optimal solution X , we go through every element of X , if the value is larger than a threshold t , we can take these regions of scene image as the final matching result.

2.3 Result Optimization

Hyper-graph matching algorithm describes the whole structure of object by high-order topology. The association among points (or regions) reduces the possibility of mismatch. However, because of the association, wrong matching points (or regions) which are retained will reduce the similarity of right matching points (or regions), it will lead to the difficulty of distinguishing them, even the similarity of wrong one is higher than the similarity of right one (Fig. 2 group(1)). To solve the problem, we will run our algorithm twice. In the first iteration, we set a low threshold t to get more matching regions. In the second iteration, we relax constraints $X \in \{0, 1\}^{n_m \times n_s}, X 1_{n_s} = 1_{n_m}, X^T 1_{n_m} = 1_{n_s}$ to

$$X \in [0, 1]^{n_m \times n_s}, X^T 1_{n_m} \leq 1_{n_s}, X 1_{n_s} \leq 1_{n_m}, 1_{n_m}^T X 1_{n_s} = u' \quad (10)$$

where u' is the expected number of matching regions, generally, $u' \approx n_m$. $X 1_{n_s} = 1_{n_m}$ in Eqn. (8) denotes all regions in model image should be matched into the scene, $1_{n_m}^T X 1_{n_s}$ in Eqn. (10) denotes almost u' regions in scene image should be matched. When some regions in model image are sheltered in scene image, Eqn. (8) will find the most similar regions to match them, obviously the results are wrong. In contrast, Eqn. (10) can assign redundant weights to other regions which are more likely to be correct results. Therefore, Eqn. (10) has better performance to resist partial shelter than Eqn. (8). However, because in the first iteration regions of scene image are often too many and the approximate value of $1_{n_m}^T X 1_{n_s}$ is uncertain, we will not use Eqn. (10) in the first iteration. Because the number of matching regions decreases rapidly, the second iteration is efficient and more iterations are needless. Besides, in order to resist partial shelter, in the second iteration we only use regions which are matched after the first iteration in model image. The performance of result optimization will be verified by experiment in Sect. 3.2.

3 Experiments

We use MSER [17] to get stable regions and remove improper regions which contain many other regions or only some pixels. We simply use RGB information (each mean value of red, green and blue in a region) as bundled feature in region term. We use RGB information or area (the number of pixels in a region) or the coordinate of MSER center as bundled feature in relation term. We use MATLAB R2011b with VLFeat open source library [18] to implement our algorithm. Typically, on Intel(R) Core(TM) i3-2120 CPU and 4.00GB RAM, for matching 8 regions in model image and 177 regions in scene image, our algorithm takes 14.476 seconds in the first iteration and takes 0.056 seconds in the second iteration. Time cost of our result optimization is little.

Generally, based on our algorithm without optimization, the weight parameters (λ, B) are 1, threshold t is 0.5. However, the values often need to be adjusted. With result optimization, the values of parameters are more stable, the weight parameters (λ, B) are 1, threshold t is 0.1 in the first iteration and is 0.5 in the second iteration.

3.1 Object Matching Experiment

Star topology of our bundled feature can describe more relationship, it can resist linear variance. Besides, features extracted from region are not sensitive to local deformation. Therefore, in theory, our algorithm has good performance on non-rigid object matching. In order to verify it, we mainly focus on matching patterns on cloth which often deforms greatly. Samples of many typical datasets are not suitable for our experiment. Therefore, we use some real images of cloth as samples. Fig. 1 shows experimental results based on our algorithm (without optimization) and method in [7, 8] (we use SIFT code provided by VLFeat library [18] and matching code provided by [7, 8] to realize the program). The biggest blue cube in Fig. 1 group(1) are man-made noise. Group(1)(2) verify the performance on rigid object and group(3-6) verify the performance on non-rigid object. In Fig. 1, most of results based on our algorithm are right. However, based on the method in [7, 8], results of Fig. 1 group(1)(2)(4)(5) are unsatisfactory, there are many wrong matching points, especially group(4)(5). By comparing the experimental results, our algorithm outperforms the method in [7, 8].

3.2 Result Optimization

In order to verify the performance of our optimization, we take a typical sample to illustrate. In Fig. 2 group(1), after the first iteration of our algorithm, we can get the matching result (Fig. 2 group(1) (c)): the similarity of region $a = 0.627, b = 0.526, c = 0.671, d = 0.543, e = 1, f = 0.556, g = 0.461, h = 0.526$. a, b and c are wrong matching regions, because of the problem mentioned in Sect. 2.3, the similarity of wrong matching region c is higher than most of right matching regions. With our optimization, we can get the matching result (Fig. 2 group(1) (d)): regions d (the similarity is 0.856), e (the similarity is 0.852),

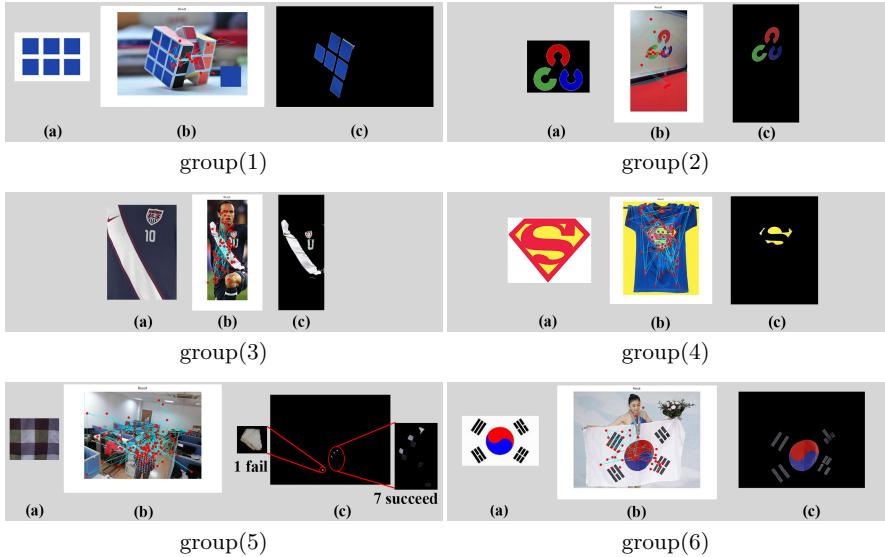


Fig. 1. Matching experiment based on our algorithm (without optimization) and method in [7, 8]. In each group, (a) is model image, (b) is scene image (without red points and blue lines), (c) is matching result based on our algorithm. In (b), red points are matching results based on method in [7, 8], blue lines are their geometric structure.

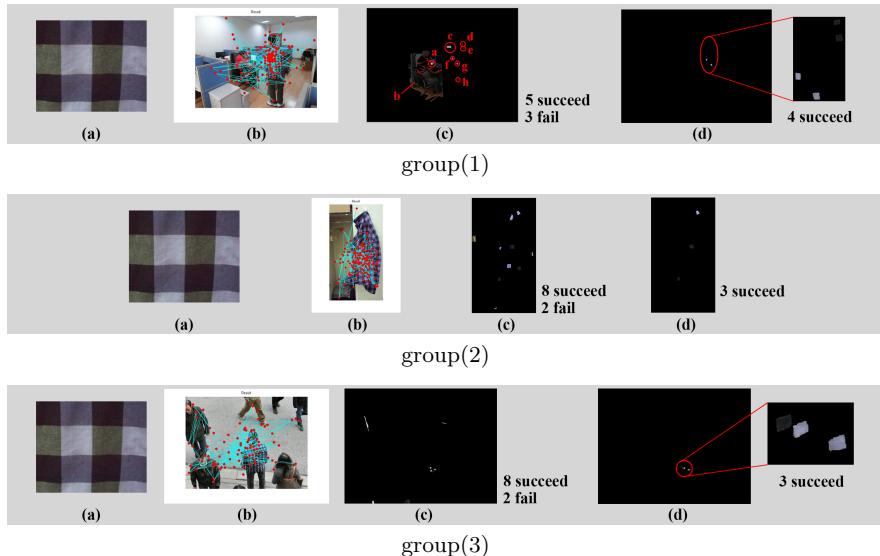


Fig. 2. Matching result Optimization. (a) is model image, (b) is scene image (without red points and blue lines), (c) is the matching result after the first iteration based on our algorithm, (d) is the matching result after the second iteration based on our algorithm. In (b), red points are matching results based on method in [7, 8], blue lines are their geometric structure.



Fig. 3. Matching result Optimization of Fig. 1 group(5). (a) is model image, (b) is scene image (without red points and blue lines), (c) is the matching result based on our algorithm with optimization. In (b), red points are matching results based on method in [7, 8], blue lines are their geometric structure.

f (the similarity is 0.524) and g (the similarity is 0.727) are retained, they are all right matching regions. The highest similarity of wrong regions is 0.419 (region a), it is easy to distinguish. Fig. 2 group(2)(3) also can verify the good performance of result optimization. With our optimization, the wrong matching region in Fig. 1 group(5) also can be removed (Fig. 3).

4 Conclusion

This paper presents an object matching algorithm based on hyper-graph method: we combine several kinds of features in stable region of object as bundled feature, it can describe more relationship by hyper-graph model, not just physical location. By introducing coefficients and definitions, we properly extend the approach in [1, 7, 8] to build and solve the hyper-graph model. Finally, we optimize the matching result by iteration and modifying constraints, it improves the accuracy effectively. By comparing to the algorithm in [7, 8], our proposed algorithm shows good performance on experiments, especially for non-rigid object matching. However, in order to make our algorithm more practical, our future work will concentrate on matching the object which cannot be divided into several proper regions. In this case, the performance of our algorithm will be affected.

Acknowledgement. This research is sponsored by the National Science Foundation of China under Grant No.61075012.

References

1. Jiang, H., Drew, M.S., Li, Z.-N.: Matching by linear programming and successive convexification. *IEEE Trans. PAMI* 29, 959–975 (2005)
2. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proc. CVPR, pp. 2161–2168 (2006)
3. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. *IEEE Trans. PAMI* 19, 530–535 (1997)
4. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proc. ICPR, vol. 1, pp. 582–585 (1994)

5. 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)
6. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: Proc. ICCV, pp. 2564–2571 (2011)
7. Li, H., Kim, E., Huang, X., He, L.: Object matching with a locally affine-invariant constraint. In: Proc. CVPR, pp. 1641–1648 (2010)
8. Li, H., Huang, X., He, L.: Object matching using a locally affine invariant and linear programming techniques. *IEEE Trans. PAMI* 35(2), 411–424 (2013)
9. Xu, L., Zhou, Y., Qingshan, L.: Region based on object recognition in 3d scenes. In: Proc. IVS, pp. 13–16 (2011)
10. Zass, R., Shashua, A.: Probabilistic graph and hypergraph matching. In: Proc. CVPR, pp. 1–8 (2008)
11. Duchenne, O., Bach, F., Kweon, I., Ponce, J.: A tensor-based algorithm for high-order graph matching. In: Proc. CVPR, pp. 1980–1987 (2009)
12. Lee, J., Cho, M., Lee, K.M.: Hypergraph matching via reweighted random walks. In: Proc. CVPR, pp. 1633–1640 (2011)
13. Kannala, J., Rahtu, E., Brandt, S.S., Heikkila, J.: Object recognition and segmentation by non-rigid quasidense matching. In: Proc. CVPR, pp. 1–8 (2008)
14. Bai, X., Wang, X., Latecki, L.J., Liu, W., Tu, Z.: Active skeleton for non-rigid object detection. In: Proc. ICCV, pp. 575–582 (2009)
15. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling feature for large scale partial-duplicate web image search. In: Proc. CVPR, pp. 25–32 (2009)
16. Lowe, D.G.: Object recognition from local scaleinvariant features. In: Proc. ICCV, vol. 2, pp. 1150–1157 (1999)
17. Donoser, M., Bischof, H.: Efficient maximally stable external region (MSER) tracking. In: Proc. CVPR, vol. 1, pp. 553–560 (2006)
18. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), <http://www.vlfeat.org/>

Methods for Photomosaic Generation Based on Different Image Similarity and Division Strategies

Daqing Chang¹, Changshui Zhang¹, and Shifeng Weng^{2,*}

¹ State Key Laboratory on Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Automation, Tsinghua University, Beijing, 100084, P.R. China

² School of Electronics and Information,
Zhejiang Wanli University, Zhejiang, 315000, P.R. China

Abstract. Photomosaic is an image that is composed of small images called tiles. In our work, different measurements of similarity between tiles in the desired original image and alternative tiles in the database are used in the generation of photomosaic after classifying tiles in the original image into tiles of purity and edge. And a heuristic position searching strategy is proposed to keep the original length-width ratio of each tile in the photomosaic generated unchanged if tiles in the database have different sizes.

Keywords: photomosaic, image classification, image similarity.

1 Introduction

Photomosaic is a kind of mosaic techniques [1] and a photomosaic is an image constituted by small images selected in a database of small images called tiles, as shown in Fig 1. When viewed at a distance, the details of each tile vanish and we will see the desired original image. Elements that affect the effectiveness of algorithms for generating photomosaic is studied in [2]. The impact of the similarity between tiles, granularity of tiles in the original image and the variety of tiles in the database is compared.

The technique to generate a photomosaic with computer is first studied by Robert Silver[3]. In his method, the original image is firstly divided with grid into tiles of the same size. Then photomosaic is generated by replacing each tile with the most similar tile alternative in the database.

From the procedure above, we can see that for each tile to be matched, every tile in the database is checked whether it is the most similar one which is time

* This work is supported by NSFC (No. 61021063), 973 Program (2009CB320602), Beijing Municipal Education Commission Science and Technology Development Plan key project under (No. KZ201210005007) and Zhejiang Provincial Natural Science Foundation of China (No. Y1110661).



Fig. 1. An example of photomosaic generated by our method. The original image is scaled on the top left.

consuming if the capacity of database is big. G.D.Biasi uses a data structure of antipole tree to speed up this procedure[4].

Essentially, the process of generating photomosaic can be taken as an optimization problem, variables of which are indexes of tiles chosen at each position in the original image and the objective function to be minimized is the sum of matching error. Methods of evolutionary computing such as genetic algorithm(GA) are applied to solve the problem [5][6][7].

However, generally among algorithms exist, a uniform measurement of similarity is used to find the most similar tile in the database for each tile in the original image. It means that for each tile features used in matching are the same. However, the criterion we use when we judge with our eyes whether a tile is similar to another may vary between different tiles. For example, when matching a tile which color changes little, perhaps the most important factor we consider is whether their main color is the same. While if a tile is part of the main outline of the original image, the similarity of shape may be considered as well as the factor of color. Therefore, in our method, tiles in the original image are classified and different measurements of similarity are used for each class. And results we get usually have a good visual effect.

Furthermore, we can see that each tile in the photomosaic generated is usually scaled to the same size of the grid used to divided the original image in existing methods. Nevertheless the length-width ratio of each tile in the database may differ from the grid and the deformation may affect the content of the tile to some extent. To avoid this,we try to describe a strategy of division to keep the length-width ratio of tiles used in the photomosaic unchanged.

The rest of this paper is organized as follows. In section 2, we describe the measurement of similarity we used. Our method for photomosaic generation with classification of tiles in the original image is described in section 3. The method that keep the length-width ratio of each tile unchanged is shown in section 4. And we conclude our work in section 5.

2 Measurement of Similarity

The procedure of generating a photomosaic can been simplified as finding the most similar tile in the database to replace each tile in the original image. The importance of the measurement of similarity during this procedure is obvious. However, different kinds of similarity can be defined from different views. Intuitively, if the color distribution of two tiles is similar at corresponding position, we may probably feel that they are similar. And if the two tiles have similar shapes which means similar edge distribution, they can also been think as similar. Besides, the content of tiles can also be taken as a factor of similarity. Therefore, both the distribution of color and edge are considered in our measurement of similarity. However, as each tile in the original image is only a small part of the whole image which usually has no meaningful content, the content of the tile is not included.

2.1 Color Similarity

Let P_1 be a tile in the original image and P_2 be a tile in the database. Then color Similarity between P_1 and P_2 is formulated in (1).

$$D_{col} = \sum_{R,G,B} \sum_{i=1}^m \sum_{j=1}^n (A_1(i,j) - A_2(i,j))^2 \quad (1)$$

As there are RGB channels for a color image, the color similarity is D_{col} defined as the sum of L-2 norm distance of each channel and $A_1(i,j)$ and $A_2(i,j)$ represent the pixel value of corresponding position in each channel. Besides, as tiles in the original image and tiles in the database are not have to be of the same size. Both tiles being matched are resized to an equal size $m \times n$ before matching.

2.2 Edge Similarity

Just like the color similarity above, the edge similarity between P_1 and P_2 is formulated in (2)

$$D_{edge} = \sum_{i=1}^m \sum_{j=1}^n (E_1(i,j) - E_2(i,j))^2 \quad (2)$$

In this formulation, D_{edge} is the edge similarity we get. E_1 is the edge strength of P_1 and E_2 is the edge strength of P_2 computed by (3) where E_h and E_v represent the horizontal gradient and vertical gradient which are the responses of mask M_1 and M_2 in (4)

$$E(i,j) = \sqrt{E_h(i,j)^2 + E_v(i,j)^2} \quad (3)$$

$$M_1 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} M_2 = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (4)$$

We can see from above that the measurement of similarity can be seen as some distance between different tiles. The smaller the distance is, the more similar the two tiles compared are.

3 Generation with Classification Of Tiles

As mentioned in section 1, tiles in the original image may have different types. For example,a tile which is part of sky in the original image may have no notable edge and a tile with the main color blue could be taken as a similar tile in the database while for a tile on the edge of some object, we prefer a similar tile with similar edge as well as similar color distribution.Therefore, to improve the visual effect of photomosaic, classification of tiles in the original image and apply different measurements of similarity to each class is used here.We propose a method to generate photomosaic with classification of tiles in the original image in this section.

3.1 Classification of Tiles in the Original Image

In section 2, we describe the measurement of similarity with color and edge considered. Just as the discussion at the beginning of this section, when matching a tile in the original image, we may focus on the factor of color for a tile smoothly changed, and perhaps both factors of color and edge are considered for a tile remarkably changed.So the classifier is trained according to the degree of change. The features we use in training are histogram, Tamura contrast and variance. As there are RGB channels for a color tile, in fact these three features are extracted in each channel and a nine-dimension vector of feature will be used for each tile in the training set.

Histogram. Histogram of each channel represents the color distribution of this tile. The denser the histogram is,the more likely that this tile is smoothly changed. In our method, the density of histogram is formulated as (5)

$$D_{hist} = \frac{\sum_{i=a}^b H(i)}{\sum_{j=1}^N H(j)} \quad (5)$$

where

$$\begin{aligned} a &= \max(1, i_{peak} - 0.1 \times N) \\ b &= \min(N, i_{peak} + 0.1 \times N) \end{aligned}$$

$H(i)$ is the value of the i -th grayscale level in histogram of some channel. i_{peak} is the index with maximum value in $H(i)$. N is the number of grayscale level.

Variance. Since variance reflects the divergency of a group of data. We take it as a measurement of changing degree too. The variance is computed by (6)

$$\sigma^2 = \sum_{i=1}^N (i - \mu)^2 h(i) \quad (6)$$

where μ is the mean value of this channel and $h(i)$ is the i-th grayscale level in histogram normalized by its maximum value.

Tamura Contrast. Contrast measures the change of brightness which is also an aspect of changing degree in a tile. In our method we exploit the definition of contrast in [8] which is formulated in (7).

$$F_{con} = \frac{\sigma}{(\alpha_4)^{1/4}} \quad \alpha_4 = \frac{\mu_4}{\sigma^4} \quad (7)$$

where F_{con} is the contrast called Tamura contrast. σ is the standard deviation which is the root of variance and μ_4 is the fourth-order moment of this channel defined as

$$\mu_4 = \sum_{i=1}^N (i - \mu)^4 h(i)$$

We randomly pick out thousands of tiles divided from thousands of original images randomly chosen as the training set. Then, a clustering algorithm of K-means is applied to label the training set. After that the classifier is trained by SVM with this labeled data.

3.2 Tile Similarity

In the last part, we have trained the classifier that used to judge whether a tile belongs to a tile of purity that means smoothly changed or to a tile of edge that means remarkably changed. Let P_1 be a tile in the original image and P_2 be a tile in the database. Then the similarity between P_1 and P_2 is formulated in (8)

$$D = \begin{cases} D_{col} & P_1 \in \text{purity} \\ (1 - \alpha)D_{col} + \alpha D_{edge} & P_1 \in \text{edge} \end{cases} \quad (8)$$

where D is the measurement of similarity. D_{col} and D_{edge} are the similarity of color and edge described in section 2. α is the weight that make balance of the two parts. Generally, we take the value of α as the L-2 norm of the edge strength of P_1 defined in (3) normalized by the maximum value that edge strength may take at a position. We know from this that for a tile belonging to tile of edge, the weight of edge similarity increases with the grow of edge strength which is obviously reasonable.

3.3 Algorithm

After defining the similarity between tiles, we describe the algorithm of our method here and Fig 2 shows some results of our method.

- 1 Divide the original image into tiles of the same size with grid.
- 2 classify each of the tiles into tile of purity and tile of edge with classifier trained before.
- 3 For each tile in the original image,compute the similarity with each tile in the database and take the one with the smallest value as the most similar tile in the database.
- 4 Replace the tiles in the original image with tiles found in the last step to achieve the photomosaic.

4 Generation without Equal Division

It is easy to see that a common characteristic of "traditional" photomosaic is that every tile in the photomosaic has the same size. However, tiles in the database perhaps have different sizes when we get them. Therefore, the length-width ratio of them maybe changed when we resize them to the fixed size used in photomosaic which will affect the content of them to some extent.Because of this, we describe a method to generate photomosaic assembled with tiles of variant sizes in this section.

4.1 Position Searching Strategy

As we want tiles appeared in the photomosaic to keep the original length-width ratio unchanged, the position of tiles cannot be decided in advance like method in the last section by dividing the original image into equal tiles with gird. Therefore, we should decide the position that the next tile to be put at before finding the most similar tile in the database.

The bad thing that comes with the variety of sizes and arbitrariness of position is overlap. We can see that the main situation that overlap may probably be produced is that we have to fill narrow gaps or small regions unmatched with tiles which are bigger than them. Hence, our target is to decrease the generation of narrow gaps and small regions unmatched. Under this criterion, our strategy is that each time we take one of the corners of region that has not been matched as the position where we put the next tile in the database. This procedure is explained in Fig 3.

4.2 Tile Similarity

After the position of the next tile to be placed at is determined, let P_1 be the tile in the original image with the same size of the tile to be measured in the database

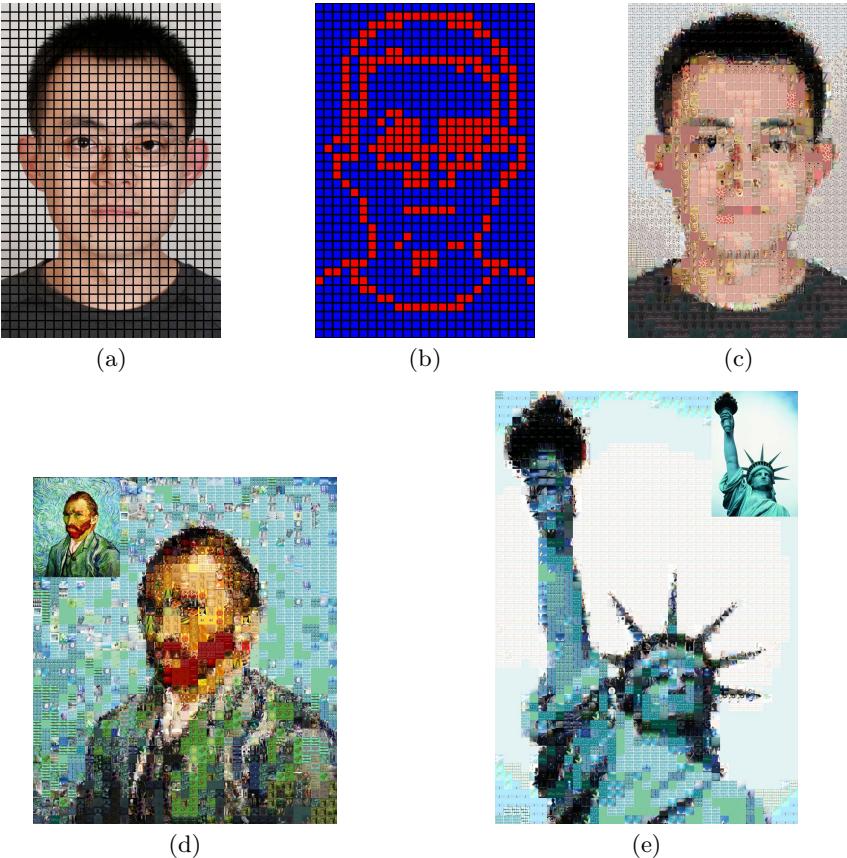


Fig. 2. Results of our method in section 3. In (a), The original image is divided in to 38×25 tiles. (b) is the result of classification that red blocks mean tiles of edge and blue blocks mean tiles of purity. (c) is the photomosaic generated. The other two results are shown in (d) and (e). The capacity of database is 5000.

at the chosen position and P_2 be the tile in the database. The measurement of similarity between P_1 and P_2 is shown in (9).

$$D = D_{col} + D_{var} + D_{shape} \quad (9)$$

where D represents the similarity we get, D_{col} is the color similarity defined in section 2. D_{var} is the sum of variances in each channel of P_1 . If P_1 is complex and sharply changed, this term is big and a smaller tile will be chosen, as the smaller the tile is, the less details there are in this tile which means the variance is relatively small. Otherwise, if P_1 is smoothly changed, this term is small and replacement with a bigger tile is permitted. D_{shape} is proportion of overlap with region that has been matched and the area out of the border of the original image in the whole area of P_1 which is a penalty for overlap. The smaller this term is, the less overlap happens.

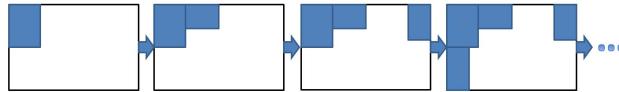


Fig. 3. An explain of position searching strategy,each tile is put at one of the corners in the unmatched region

4.3 Algorithm

At last of this section, we describe the complete algorithm as follows and results of our algorithm are shown in Fig 4:

- 1 Find the next position to be matched next according to the position searching strategy.
- 2 Compute the similarity between the tile at the chosen position in last step and tiles in the database and replace the tile in the original image with the most similar tile in the database.
- 3 Repeat the first two steps until all regions in the original image are replaced.

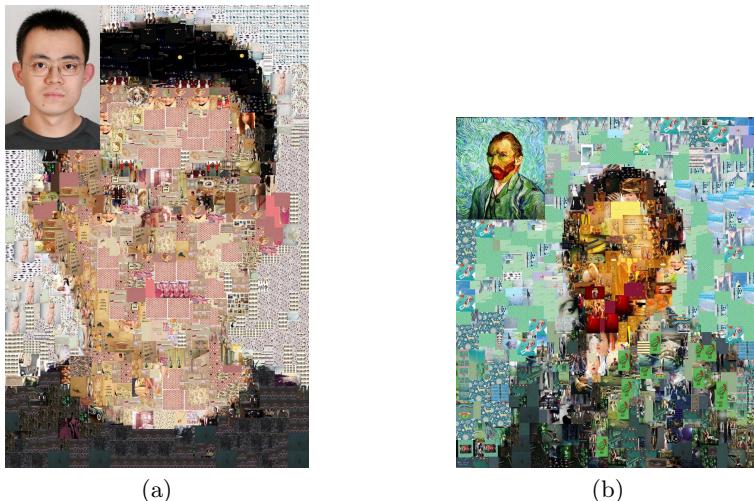


Fig. 4. Results of method in section 4. 1006 tiles are used in (a) and 839 tiles are used in (b). The capacity of database is 5000.

5 Conclusion

In our work, we propose a method for photomosaic generation with classification of tiles in the original image and different measurements of similarity are applied to find the best tile in the database. Besides, we also describe a method to generate photomosaic with tiles of arbitrary sizes. However, we don't classify

tiles in the original image here as in the first case. This is because the features we use to train the classifier don't have the character of invariance with the change of sizes and how to deal with the problem is one of the work we need to do in the future.

Furthermore, another thing we will do next is reducing the repeat of a single tile in the photomosaic. If we restrict the maximum times that a single tile can be used in the photomosaic, we can't get the best photomosaic of the minimum sum of error by matching each tile line-by-line, which is more evident if the capacity of the database is not big enough.

Also, how to speed up the procedure of generation will be considered in our future work.

References

1. Battiatto, S., Di Blasi, G., Farinella, G.M., Gallo, G.: Digital mosaic frameworks - an overview. *Computer Graphics Forum* 26(4), 794–812 (2007)
2. Tran, N.: Generating photomosaics: an empirical study. In: *Proceedings of the 1999 ACM Symposium on Applied Computing*, SAC 1999, pp. 105–109. ACM, New York (1999)
3. Silvers, R.: Photomosaics. Henry Holt and Co., Inc., New York (1997)
4. Di Blasi, G., Petralia, M.: Fast photomosaic (2009)
5. Mat Sah, S.B., Ciesielski, V., D'Souza, D., Berry, M.: Comparison between genetic algorithm and genetic programming performance for photomosaic generation. In: Li, X., Kirley, M., Zhang, M., Green, D., Ciesielski, V., Abbass, H.A., Michalewicz, Z., Hendtlass, T., Deb, K., Tan, K.C., Branke, J., Shi, Y. (eds.) *SEAL 2008*. LNCS, vol. 5361, pp. 259–268. Springer, Heidelberg (2008)
6. Mat Sah, S.B., Ciesielski, V., D'Souza, D.: Refinement techniques for animated evolutionary photomosaics using limited tile collections. In: Di Chio, C., Brabazon, A., Di Caro, G.A., Ebner, M., Farooq, M., Fink, A., Grahl, J., Greenfield, G., Machado, P., O'Neill, M., Tarantino, E., Urquhart, N. (eds.) *EvoApplications 2010*, Part II. LNCS, vol. 6025, pp. 281–290. Springer, Heidelberg (2010)
7. Narasimhan, H., Satheesh, S.: A randomized iterative improvement algorithm for photomosaic generation. In: *World Congress on Nature Biologically Inspired Computing*, NaBIC 2009, pp. 777–781 (December 2009)
8. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics* 8(6), 460–473 (1978)

Robust Continuous Terminal Sliding Mode Control Design for a Near-Space Hypersonic Vehicle

Ruimin Zhang^{1,2,*}, Changyin Sun¹, Jingmei Zhang¹, and Chengshan Qian¹

¹ School of Automation, Southeast University, Nanjing, 210096, China

² School of Mathematics & Statistics, Henan University of Science & Technology, Luoyang, 471003, China

Abstract. This paper presents a robust continuous terminal sliding mode attitude tracking control approach for a near-space hypersonic vehicle(NSHV) in the presence of parameter uncertainties and external disturbances. Firstly, a novel nonsingular terminal sliding surface is developed. Then, a continuous terminal sliding mode control law is proposed, which is chattering-free. Afterward, considering the presence of parameter uncertainties and external disturbances, a high order sliding mode disturbance observer is introduced to estimate the lumped disturbance to improve the control performance. Finally, numerical simulations applied to the NSHV illustrate the effectiveness of the proposed approach.

Keywords: Terminal sliding mode control, Finite-time convergence, Near-space hypersonic vehicle, Tracking control.

1 Introduction

NSHVs offer a great potential for feasible access to space and high speed civil transportation. The control design for NSHVs faces significant challenging because of NSHVs' strong nonlinearity, highly time-varying dynamics, large parameter uncertainties. Up to now, various nonlinear control approaches have been presented to tackle this problem, such as dynamic inversion [1], backstepping control[2,3], sliding mode control (SMC) [4,5].

Among these control approaches, sliding mode control is a well-known powerful control scheme which has many attractive characteristics such as good transient, fast response and insensitivity to parameter uncertainties and external disturbances. Therefore, SMC has been widely applied for flight control. However, SMC has two disadvantages. The first is that the linear surface usually used in SMC design only can guarantee the asymptotic stability. The second is chattering phenomena, which is the main drawback of SMC.

* This work is supported by National Outstanding Youth Science Foundation (61125306); National Natural Science Foundation of Major Research Plan (91016004, 61034002), Specialized Research Fund for the Doctoral Program of Higher Education of China (20110092110020).

To address these issues, in this study, we develop a new robust continuous terminal sliding mode attitude tracking control scheme for the NSHV under the parameter uncertainties and external disturbances. Firstly, a novel nonsingular terminal sliding surface is proposed. Then, based on the Lyapunov method, a continuous terminal sliding mode control law is derived to guarantee the existence of the sliding mode within finite time. Further, to cope with the parameter uncertainties and external disturbances, a high order sliding mode disturbance observer is used to estimate the lumped disturbance. Finally, the proposed approach is applied to design the attitude control system for the NSHV and simulation results is presented to demonstrate the effectiveness of the current method.

2 Robust Continuous Terminal Sliding Mode Control Design

A nonlinear system under model uncertainties and external disturbances is described as

$$\dot{x} = f(x) + \Delta f(x) + (g(x) + \Delta g(x))u + d(t) \quad (1)$$

where $x \in \mathbb{R}^n$ is the state vector of the system. $u \in \mathbb{R}^n$ is the control input vector. $f(x) \in \mathbb{R}^n$, $g(x) \in \mathbb{R}^{n \times n}$ are known nonlinear system functions of the state variables and time. Moreover, $g(x)$ is invertible. $\Delta f(x)$, $\Delta g(x)$ represent model uncertainties. $d(t)$ denotes the external disturbance. The lumped disturbance $\psi = \Delta f + \Delta g u + d$ represents a lumped disturbance, which is differentiable and has a known Lipschitz constant $L_i > 0$.

To design nonsingular terminal sliding mode control, a novel nonsingular terminal sliding surface is designed as:

$$s = e + \int_0^t (k_1 \text{sig}^{\gamma_1}(e) + k_2 \text{sig}^{\gamma_2}(e))d\tau \quad (2)$$

where $e = x - x_d$, $k_1 = \text{diag}(k_{11}, \dots, k_{1n})$, $k_2 = \text{diag}(k_{21}, \dots, k_{2n})$. $k_{ij} > 0$ ($i = 1, 2, j = 1, \dots, n$), $\gamma_1 \geq 1$ and $0 < \gamma_2 < 1$ are constants.

Once the tracking error reaches the sliding surface, it satisfies the equation $\dot{s} = 0$. Then the sliding mode dynamics is derived as follows:

$$\dot{e} = -k_1 \text{sig}^{\gamma_1}(e) - k_2 \text{sig}^{\gamma_2}(e) \quad (3)$$

By solving the differential equation (3), it can be obtained that $e_{si} = 0$ will be reached in a finite time determined by

$$t_{si} = \int_0^{|e_{si}(0)|} \frac{1}{k_{s1} e^{\gamma_{s1}} + k_{s2} e^{\gamma_{s2}}} de_{si} = \frac{|e_{si}(0)|^{1-\gamma_{s1}}}{1-\gamma_{s1}} k_{s1}^{(1-\gamma_{s1})/\gamma_{s1}} \times F\left(1, \frac{\gamma_{s1}-1}{\gamma_{s1}-\gamma_{s2}}; \frac{2\gamma_{s1}-\gamma_{s2}-1}{\gamma_{s1}-\gamma_{s2}}; -k_{s2} k_{s1}^{-1} \|e_{si}(0)\|^{\gamma_{s2}-\gamma_{s1}}\right), i = 1, 2, 3 \quad (4)$$

where $F(\cdot)$ denotes Gauss' Hypergeometric function.

After the nonsingular terminal sliding surface is established, then a nonsingular terminal sliding mode controller is proposed without considering the presence of parameter uncertainties and external disturbances, as follows:

$$u = -g^{-1}[f - \dot{x}_d + k_1 \text{sig}^{\gamma_1}(e) + k_2 \text{sig}^{\gamma_2}(e) - l_1 s + l_2 \text{sig}^{\eta}(s)] \quad (5)$$

where $l_1 = \text{diag}(l_{11}, \dots, l_{1n})$, $l_2 = \text{diag}(l_{21}, \dots, l_{2n})$, $0 < \eta < 1$. η and l_{ij} ($i = 1, 2, j = 1, \dots, n$) are positive constants.

Theorem 1. Considering the nonlinear system(1) in the absence of parameter uncertainties and external disturbances, if the sliding surface is designed as (2) and the controller is constructed as (5), then the tracking error e will reach the sliding surface in a finite time T , given by

$$T \leq \frac{\ln(\frac{l_2 - 2^{\frac{1-\eta}{2}} \bar{l}_1 V^{\frac{1-\eta}{2}}}{\underline{l}_2})}{\bar{l}_1(1-\eta)} \quad (6)$$

Proof: Differentiating sliding variable (2) and using (1) and (5), we can get the closed-loop sliding dynamic equation as

$$\dot{s} = l_1 s - l_2 \text{sig}^\eta(s) \quad (7)$$

Consider the following Lyapunov function candidate $V = \frac{1}{2}s^T s$

Taking the time derivative of Lyapunov function and using (7), one can get

$$\dot{V} = s^T \dot{s} = s^T [l_1 s - l_2 \text{sig}^\eta(s)] = s^T l_1 s - s^T l_2 \text{sig}^\eta(s) \quad (8)$$

$$\leq \bar{l}_1 \|s\|^2 - \underline{l}_2 \|s\|^{\eta+1} = 2\bar{l}_1 V - 2^{\frac{\eta+1}{2}} \underline{l}_2 V^{\frac{\eta+1}{2}} \quad (9)$$

where $\bar{l}_1 = \max_{i=1, \dots, n} \{l_{1i}\} > 0$ and $\underline{l}_2 = \min_{i=1, \dots, n} \{l_{2i}\} > 0$.

Therefore, according to Lemma 2 in [7] , the tracking error $e(t)$ will reach the sliding surface in a finite time T .

Once the sliding mode $s = 0$ is reached, tracking error $e(t)$ will converge to zero in the sliding mode within a finite time .

Remark 1. The terminal sliding mode control given in (5) is a novel continuous second order sliding mode control since the condition $s = \dot{s} = 0$ is satisfied on the sliding surface[6].

When considering the presence of parameter uncertainties and external disturbances of nonlinear system (1), the upper bound of the lumped disturbance ψ is usually unknown. Therefore, a high order sliding mode disturbance observer is introduced to estimate the lumped disturbance. ψ_i can be smoothly estimated by the following high order siding mode disturbance[6].

$$\begin{aligned} \dot{z}_{0i} &= v_{0i} + u_i \\ v_{0i} &= -2L_i^{1/3} |z_{0i} - s_i|^{2/3} \text{sign}(z_{0i} - s_i) + z_{1i} \\ z_{1i} &= v_{1i} \\ v_{1i} &= -1.5L_i^{1/2} |z_{1i} - v_{0i}|^{1/2} \text{sign}(z_{1i} - v_{0i}) + z_{2i} \\ z_{2i} &= -1.1L_i \text{sign}(z_{2i} - v_{1i}) \\ i &= 1, \dots, n \end{aligned} \quad (10)$$

Then $z_1 = [z_{11}, z_{12}, \dots, z_{1n}]^T = \hat{\psi}$ converges to ψ in a finite time, if the sliding variable s and control input u are measured without noise.

Finally, based on the combination of (5) and (10), a robust continuous terminal sliding mode control (RCTSMC) for the nonlinear system(1) is proposed under parameter uncertainties and external disturbances.

$$u = -g^{-1}[f - \dot{x}_d + k_1 \text{sig}^{\gamma_1}(e) + k_2 \text{sig}^{\gamma_2}(e) - l_1 s + l_2 \text{sig}^\eta(s) + z_1] \quad (11)$$

where $l_1 = \text{diag}(l_{11}, \dots, l_{1n})$, $l_2 = \text{diag}(l_{21}, \dots, l_{2n})$, $0 < \eta < 1$. η and l_{ij} ($i = 1, 2, j = 1, \dots, n$) are positive constants.

Differentiating sliding variable (2) and using (1) and (11), we can obtain the following closed-loop dynamics

$$\dot{s} = l_1 s - l_2 \text{sig}^\eta(s) - z_1 + \phi \quad (12)$$

After z_1 converges to ψ in finite time. The system is reduced to the system

$$\dot{s} = l_1 s - l_2 \text{sig}^\eta(s) \quad (13)$$

The system (13) repeats the system (7) which is finite-time stable.

The approach described above will be used to carry out tracking simulations for the NSHV discussed in next Section.

3 Simulation Applied to the NSHV

The considered attitude control model of the NSHV is derived from the six-degree of freedom and twelve-state kinematic equations which can be simplified as the affine nonlinear equation as follows[8]:

$$\begin{cases} \dot{\Omega} = f_s + \Delta f_s + (g_s + \Delta g_s)\omega \\ \dot{\omega} = f_f + \Delta f_f + (g_f + \Delta g_f)M_c + d(t) \end{cases} \quad (14)$$

where $\Omega = [\alpha, \beta, \mu]^T$ is the state vector of the slow loop which is the attitude angle of NSHV including angle of attack, sideslip angle and bank angle. $\omega = [p, q, r]^T$ is the fast-loop state vector. M_c is the control torque vector. $g_s, g_f \in R^{3 \times 3}$ are the invertible matrices and $f_s, f_f \in R^3$. The concrete expressions of the above matrixes are specified in [8]. $\Delta f_s, \Delta f_f, \Delta g_s$ and Δg_f are model uncertainties induced by the system parameter uncertainties. $d(t)$ is the external disturbance. Next, according to Theorem 1, we will design the controller for the NSHV(14) to carry out the simulation. For comparative study, the conventional terminal sliding mode control (CTSMC) law based on the constant-rate reaching law is also employed to design the controller for the NSHV. Simulation results are shown in Fig.1.

The simulations are carried out for initial conditions with $V_0 = 2.6 km/s$, $H_0 = 30 km$, $\alpha_0 = 0.1^\circ$, $\beta_0 = 0^\circ$, $\mu_0 = 0.1^\circ$ and $p_0 = q_0 = r_0 = 0^\circ/s$. The command signals are chosen to be $\alpha_c = 1.5^\circ$, $\beta_c = 0^\circ$, $\mu_c = -1^\circ$. Assume that there exist $-30\% \sim +30\%$ random uncertainties in the aerodynamic coefficients. Besides, the external disturbance moment is defined as $d(t) = 10^4 \sin(t) [1 \ 1 \ 1]^T N \cdot M$. The parameters of the controller designed by the proposed approach are set

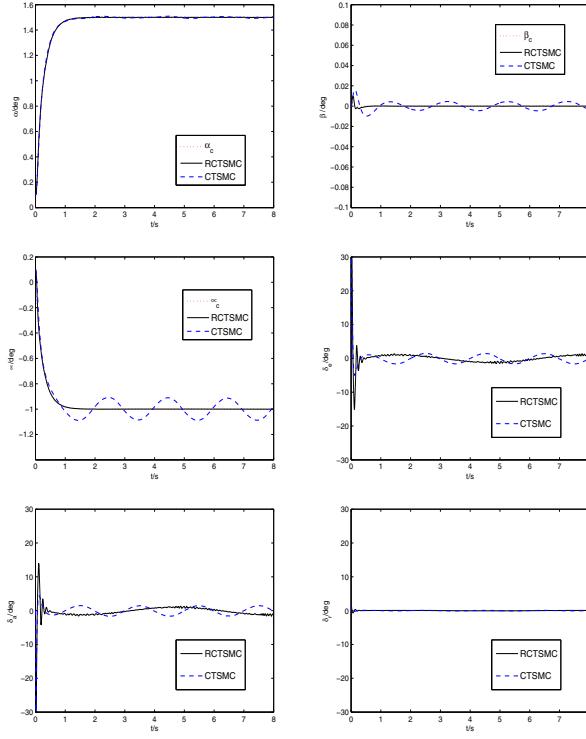


Fig. 1. Comparison of attitude control performance

as $\gamma_{s1} = 1.2$, $\gamma_{s2} = 0.8$, $k_1^s = 0.2I$, $k_2^s = 0.5I\eta_s = 0.7$, $l_1^s = 3I$, $l_2^s = 2I$, $\gamma_{f1} = 1.2$, $\gamma_{f2} = 0.8$, $k_1^f = 0.2I$, $k_2^f = 0.5I$, $\eta_f = 0.82$, $l_1^f = 1I$, $l_2^f = 4I$.

Fig.1 shows that the proposed approach attains a higher precision and better dynamic performance than CTSMC approach. Moreover, it can be referred that the proposed approach can eliminate the chattering phenomena and provide a continuous tracking control law.

4 Conclusions

In this paper, we have described the design of a robust continuous terminal sliding mode controller for the NSHV. The proposed approach not only provides a finite-time convergence, but also eliminates the chattering phenomena. Simulation experiments show good performance of the presented approach for the NSHV attitude control.

References

1. Da Costa, R.R., Chu, Q.P., Mulder, J.A.: Reentry flight controller design using nonlinear dynamic inversion. *Journal of Spacecraft and Rockets* 40(1), 64–71 (2003)
2. Li, J., Ren, Z., Qu, X.: Design of active disturbance rejection backstepping attitude controller for maneuvering glide vehicles. *Journal of Systems Engineering and Electronics* 32(8), 1711–1715 (2010)
3. Zhang, J., Sun, C., Huang, Y.: DSC-backstepping based robust adaptive LS-SVM control for Near-space vehicle's reentry attitude. *International Journal of Intelligent Computing and Cybernetics* 5(3), 381–400 (2012)
4. Ning, G., Zhang, G., Fang, Z.: Entry control using sliding modes and state observer synthesis for reusable launch vehicle. *Journal of Astronautics* 28(1), 69–76 (2007)
5. Xu, H., Mirmirani, M., Ioannou, P.A.: Adaptive sliding mode control design for a hypersonic flight vehicle. *Journal of Guidance, Control, and Dynamics* 27(5), 829–838 (2004)
6. Levant, A.: Sliding order and sliding accuracy in sliding mode control. *Int. Journal of Control* 58(6), 1247–1263 (1993)
7. Shen, Y., Xia, X.: Semi global finite time observers for nonlinear systems. *Automatica* 4(12), 3152–3156 (2008)
8. Cheng, L., Jiang, C.S., Pu, M.: Online-SVR-compensated nonlinear genealizred predictive control for hypersonic vehicles. *Science Chian Information* 54(3), 551–562 (2011)

A Level Set with Shape Priors Using Moment-Based Alignment and Locality Preserving Projections

Bin Wang¹, Xinbo Gao¹, Jie Li¹, Xuelong Li², and Dacheng Tao³

¹ School of Electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, P.R. China

² Center for OPTical IMagery Analysis and Learning (OPTIMAL),

State Key Laboratory of Transient Optics and Photonics,

Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences,

Xi'an 710119, Shaanxi, P.R. China,

³ Centre for Quantum Computation & Intelligent Systems, University of Technology,

Sydney, Australia

Abstract. A novel level set method (LSM) with shape priors is proposed to implement a shape-driven image segmentation. By using image moments, we deprive the shape priors of position, scale and angle information, consequently obtain the aligned shape priors. Considering that the shape priors sparsely distribute into the observation space, we utilize the locality preserving projections (LPP) to map them into a low dimensional subspace in which the probability distribution is predicted by using kernel density estimation. Finally, a new energy functional with shape priors is developed by combining the negative log-probability of shape priors with other data-driven energy items. We assess the proposed LSM on the synthetic, medical and natural images. The experimental results show that it is superior to the pure data-driven LSMs and the representative LSM with shape priors.

Keywords: Image segmentation, active contour model, level set, shape priors, moment-based alignment, locality preserving projections.

1 Introduction

Image segmentation is an important foundation of advanced image analysis, image understanding and pattern recognition, etc. In past decades, numerous segmentation techniques have been proposed based on different theories [1]. In recent years, active contour models (ACMs) based on partial differential equation (PDE) have been well investigated and gradually played an important role in image segmentation.

ACMs firstly were well known as the Snake algorithm [2] which is widely used in image segmentation and object recovery, etc. However owing to the explicit representation of the evolving curve, Snake is encumbered with the problem of naturally handling the topological changing of evolving curve. Therefore, Osher and Sethian [3] proposed a signed distance function (SDF) defined in higher dimensional space to describe a closed planar curve. SDF allows the evolving curve to change its topology [4] (e.g., merging and splitting) and facilitate the comparison with the different topological curves. It is using the zero level set of SDF to format the planar curve that the implicit ACMs are also called LSMs.

Following [3], a great deal of methods [1][5][6] are proposed by using different specified edge indicators to stop the curve at image edges. This kind of methods is not robust against noise and cannot effectively handle the leakage for the weak boundaries. To address this problem, Chan and Vese introduced the Mumford-Shah model [8] into LSMs [7] and boosted the region-based LSMs. [7] is more robust against noise and copes with weak boundaries better. A series of methods [9][10][11][12][13] following [7] were proposed to make some improvements, however, ignoring local features of images contributes to the difficulty to deal with the segmentation of inhomogeneous images. Additionally, both edge-based and region-based LSMs cannot handle the segmentation of broken, overlapped objects as well as the objects sharing the similar features with the image background.

To address the aforementioned problem, following the work of Chen et al. [15] Chan and Zhu [14] utilized a single shape prior to constrain curve evolution. Tsai et al. [16] executed principle component analysis (PCA) [17] on the shape priors and obtained a series of eigen shapes, i.e., eigen vectors in matrix form. Under the assumption that a given shape can be approximated by the linear combination of eigen shapes, the shape-driven energy term was designed on the Euclidean distance between the shape and its linear approximation. Instead of estimating the affine parameters, Leventon et al. [18] utilized maximum a-posteriori (MAP) to design a shape energy term after applying PCA on the aligned shape priors. Following [18], Derraz et al. [19] replaced PCA with Kernel-PCA to capture the real structure in low dimensional space. Cremers et al. [20] made an intrinsic alignment on the shape priors, and then employed kernel density estimation (KDE) to build a shape energy term. Unlike the other methods [16][21], the intrinsic alignment does not need to iteratively compute the affine parameters to minimize the energy function of alignment. However, there are still some concerns worth improving as follows. First, [20] does not provide rotation transformation to the unaligned shape priors. Second, the shape priors distributing sparsely in the observation space causes the real distribution structure to be hard estimated.

To overcome these problems, we propose a novel LSM with shape priors. Based on [20], the shape priors are aligned by using image moments. For the case with multiple shape priors, LPP [24] is utilized to reduce the dimensionality of shape priors and to capture their real statistical distribution. The proposed model has three advantages as follows. Firstly, we use the moment-based alignment to deprive the shape priors of scale, position and angle information. Secondly, the statistical distribution of the shape priors is observed in low dimensional subspace expanded by using LPP. This subspace preserves the locality relationship of shape priors in observation space, and contributes to modelling their distribution more accurately. The proposed method is compared with [7] and [20], and the results show that the proposed method obtains the competitive performance than the two methods.

The rest of this paper is organized as follows. Section 2 presents the proposed method with shape priors including two cases, i.e., single shape prior and multiple shape priors, respectively. Section 3 evaluates the performance of the proposed method comparing with other LSMs [7][20] on synthetic, natural and medical images. Section 4 is the conclusion.

2 The Proposed Method

The proposed method is two-fold. First, the shape priors are aligned by using image moments. Second, the general energy containing the shape-driven energy item, e.g., the energy for single or multiple shape priors, are minimized to implement the segmentation constrained by shape priors.

2.1 The Moment-Based Shapes Alignment

Like the most of LSMs with shape priors e.g., [14][20][18][16], our method also employs the level set function as the shape descriptor, as shown in Fig. 1. By this descriptor, the shapes with different topologies can be easily compared.

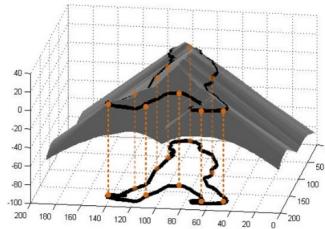


Fig. 1. The level set function as a shape descriptor

Based on the work of [20], here, rotation transformation is supported, an enhanced moment-based alignment is presented as follows.

$$\psi_i(x, y) = s_i \Phi_i \left(\frac{(x - \bar{x}_i) \cos \theta_i + (y - \bar{y}_i) \sin \theta_i}{s_i}, \frac{-(x - \bar{x}_i) \sin \theta_i + (y - \bar{y}_i) \cos \theta_i}{s_i} \right), \quad (1)$$

where Φ_i is the i th originally unaligned shape prior, ψ_i denotes the i th aligned shape prior, s_i is the scale factor to control the size of the shape prior, θ_i controls the main axis angle between the current shape and the shape prior; (\bar{x}_i, \bar{y}_i) is the centroids of the i th shape prior ψ_i . After alignment, the centroids of shape priors are moved to the center of image, and normalized to being the same scale and in the same orientation.

2.2 The Case with Single Shape Prior

In general, single shape prior exerts a crucial constraint on shape matching. The shape-driven energy term is defined as the squares of error between the current shape and the aligned shape prior. The energy term can be defined as

$$E_{shape} = \int_{\Omega} (H(\phi(x, y)) - H(\tilde{\psi}(x, y)))^2 dx dy, \quad (2)$$

where $\tilde{\psi}(x, y)$ is the shape obtained by applying a specified affine transformation to the shape prior $\psi(x, y)$, and the parameters of this transformation are determined by the current shape $\phi(x, y)$. By computing the Euler-Lagrange equation of Eq. (2), the evolution equation is obtained as

$$\frac{\partial E_{shape}}{\partial t} = 2\delta(\phi)(H(\phi) - H(\tilde{\psi})), \quad (3)$$

where ϕ is shorthand for $\phi(x, y)$ and $\tilde{\psi}$ for $\tilde{\psi}(x, y)$; $\delta(\cdot)$ is the Dirac function; t is the time step.

2.3 The Case with Multiple Shape Priors

In general, one object maybe has different shapes due to different views, meanwhile, shape priors impossibly cover all shapes to be segmented. Additionally, some saw tooth on the outline of shape priors deteriorate the performance of shape constraint. As aforementioned, we utilized LPP to obtain a low dimensional subspace in which KDE is applied to estimate the probability. The main difference from the work of Cremers et al. [20] is that the distances between the current shape and priors are computed in the expanded subspace. This shuns the curse of dimensionality and contributes to a more accurate estimation. The shape-driven energy term is defined as

$$E_{shape} = -\ln \left(\frac{1}{N} \sum_{i=1}^N \exp \left(-\frac{d^2(H(\phi)', H(\tilde{\psi})'_i)}{2\sigma^2} \right) \right), \quad (4)$$

where $H(\phi)'$ in the subspace is the projection of a given shape $H(\phi)$ in the observation space; $H(\tilde{\psi})'_i$ in the subspace is the projection of the prior obtained by applying a specified affine transformation to the i th aligned shape prior $H(\psi_i)$ and $d(\cdot)$ is a distance function which is Euclidean distance.

The evolution equation is defined as

$$\frac{\partial E_{shape}}{\partial t} = \frac{\sum_{i=1}^N b_i (H(\phi)' - H(\tilde{\psi})'_i) A^T \delta(\phi)}{2\sigma^2 \sum_{i=1}^N b_i}, \quad (5)$$

where A^T is the transformation matrix of LPP;

$$b_i = \exp \left(-d^2(H(\phi)', H(\tilde{\psi})'_i) / 2\sigma^2 \right), \quad (6)$$

σ is the kernel width. Follow the definition in [20], we fix σ^2 to be

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N \min_{j \neq i} d^2 \left(H(\tilde{\psi})_i^{'}, H(\tilde{\psi})_j^{'} \right). \quad (7)$$

Combining shape-driven energy term, i.e., Eq. (2) or Eq. (4), with the Chan-Vese method [6], the total energy functional is given as

$$E = \alpha E_{data} + (1 - \alpha) E_{shape}, \quad (8)$$

where α is a constant adjusting the effects of the data and shape energy terms, its value empirically falls into the range [0.2, 0.6]. E_{shape} could be Eq. (2) or Eq. (4), and E_{data} is borrowed from [7] and defined as

$$E_{data}(\phi) = \alpha \int_{\Omega} \delta(\phi) |\nabla \phi| dx dy + \nu \int_{\Omega} H(\phi) dx dy + \lambda_1 \int_{inside(C)} |I - c_1|^2 dx dy + \lambda_2 \int_{outside(C)} |I - c_2|^2 dx dy, \quad (9)$$

The evolution equation for Eq. (8) is obtained as

$$\frac{\partial \phi}{\partial t} = \alpha \delta(\phi) \left[\alpha \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \lambda_1 (I - c_{in})^2 + \lambda_2 (I - c_{out})^2 \right] + (1 - \alpha) \frac{\partial E_{shape}}{\partial t}, \quad (10)$$

where c_{in} and c_{out} are the inside and outside averages of the evolving curve, respectively; and $\partial E_{shape} / \partial t$ is the Eq. (3) or Eq. (5) for different case.

3 Experiments

To assess the performance of the proposed method, the comparison experiments on the different images are executed against the Chan-Vese method [7] named as M1 and the Cremer's method [20] named as M2. In each experiment, the segmentation result is outlined with yellow line, and the shape priors with magenta line.

Experiment 1 applied the three methods on a synthetic image, as shown in Fig. 2. The shape prior is not parallel with the shape in the image. The results show that M1 did not correctly detect the object as a letter "h" since it did not consider the shape prior; M2 also failed since it did not involve rotation; our method named M3 obtained desired result since it can rotate the shape prior according to the main axis of the current shape.

Experiment 2 made a comparison on the MRI image containing a human knee, as shown in Fig. 3. The desired object is the kneecap sharing the same density with the surrounding organs. The shape priors were obtained by manually segmenting from other MRI images. Without involving the shape priors, M1 failed to detect the object. M2 did not detecting the kneecap correctly but M3 did, which illustrates the distribution of the shape priors in the subspace is superior to the one in the observation space.

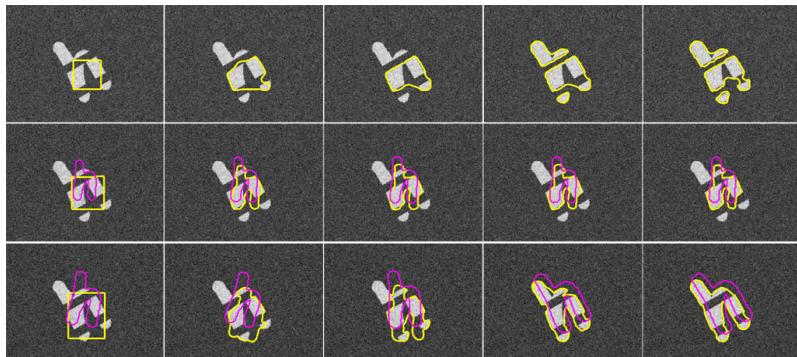


Fig. 2. The results of Experiment 1. The first to the last row represent the evolution processes and results obtained by using M1, M2 and M3, respectively.

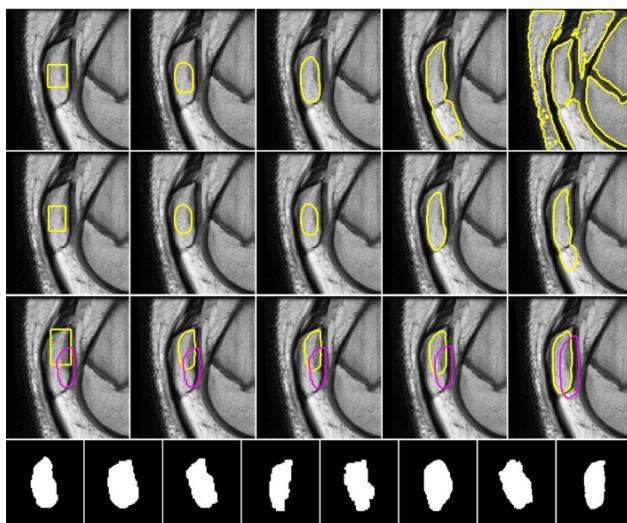


Fig. 3. The results of Experiment 2. The first to the last row represent the evolution processes obtained by using M1, M2, M3 and shape priors, respectively.

Experiment 3 illustrates the performance on a natural image with a black rabbit on grass, as shown in Fig. 4. The results show that M1 took the part of shadow as the rabbit wrongly; M2 did not include the ear of the rabbit into the curve; and M3 obtained a better result since LPP effectively preserves the nonlinear structure. This experiment suggests that the proposed method achieves better performance than M2 does on the natural images since LPP presents a more accurate statistical structure on low dimensional subspace. Additionally, the iteration number of M2 and M3 are 42 and 28, respectively.

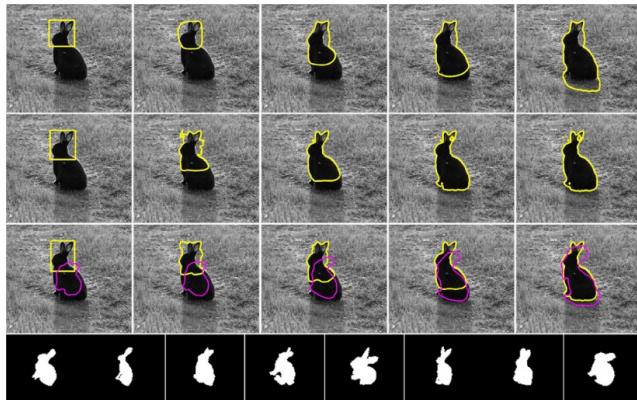


Fig. 4. The results of Experiment 3. From the first to the last row represent the evolution processes obtained by using M1, M2, M3 and shape priors, respectively.

4 Conclusion

In this paper, we proposed a novel LSM with shape priors. The proposed method firstly aligns the shape priors using image moments instead of iterative computation to simplify the procedure of alignment. Secondly, LPP reduces the dimensionality of the shape priors, meanwhile, it preserves the neighboring structure of the shape priors which contributes to a more accurate estimation. Finally, shape priors are combined with a data-driven energy term to obtain a competitive performance on broken or overlapped objects and the objects sharing common features with image background. In future, we will focus on using this energy to realize the selective segmentation.

Acknowledgement. The authors would like to thank the anonymous reviewers for their valuable comments to improve this paper. This work is partially support by the National Natural Science Foundation China under Grant 61125204 and Grant 61201293, the Fundamental Research Funds for the Central Universities under Grant K50511020016 and Grant K5051202043, the Research for the Doctoral Program of Higher Education of China under Grant 20120203120012, the China Post-Doctoral Science Foundation under Grant 20110490166.

References

1. Zhang, Y.: *Advances in Image and Video Segmentation*, PA. I RM Press, Hershey (2006)
2. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. *Int. J. Comput. Vis.* 1(4), 321–331 (1988)
3. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. *J. Comput. Phys.* 79(1), 12–49 (1988)
4. Suri, J.S., Liu, K., Singh, S., Laxminarayan, S.N., Zeng, X., Reden, L.: Shape recovery algorithms using level sets in 2D/3D medical imagery: a state-of-the-art review. *IEEE Trans. Inf. Techno. B.* 6(1), 8–28 (2002)

5. Chopp, D.L.: Computing minimal surfaces via level set curvature flow. *J. Comput. Phys.* 106(1), 77–91 (1993)
6. Malladi, R., Sethian, J.A., Vemuri, B.C.: Shape modeling with front propagation: a level set approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 17(2), 158–175 (1995)
7. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Trans. Image Process.* 10(2), 266–277 (2001)
8. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* 42(5), 577–685 (1989)
9. Chan, T.F., Sandberg, B.Y., Vese, L.A.: Active contour without edges for vector-valued images. *J. Vis. Commun. Image Represent.* 11(2), 130–141 (2000)
10. Tsai, A., Yezze Jr., A., Willsky, A.S.: Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpretation, and magnification. *IEEE Trans. Image Process.* 10(8), 1169–1186 (2001)
11. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford and Shah model. *Int. J. Comput. Vis.* 53(5), 271–293 (2002)
12. Wang, Z., Vemuri, B.C.: DTI segmentation using an information theoretic tensor dissimilarity measure. *IEEE Trans. Med. Imag.* 24(10), 1267–1277 (2005)
13. Wang, B., Gao, X., Tao, D., Li, X.: A unified tensor level set for image segmentation. *IEEE Trans. Syst., Man., Cybern. B, Cybern.* 40(3), 857–867 (2010)
14. Chan, T.F., Zhu, W.: Level set based shape prior segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., San Diego, CA, USA, vol. 2, pp. 1164–1170 (2005)
15. Chen, Y.M., Tagare, H.D., Thiruvenkadam, S., Huang, F., Wilson, D., Gopinath, K.S., Briggs, R.W., Geiser, E.A.: Using prior shapes in geometric active contours in a variational framework. *Int. J. Comput. Vis.* 50(3), 315–328 (2002)
16. Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W.E., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Trans. Med. Imaging* 22(2), 137–154 (2003)
17. Jolliffe, I.: Principal Component Analysis. Springer, New York (1986)
18. Leventon, M.E., Eric, W., Grimson, L., Faugeras, O.: Statistical shape influence in geodesic active contours. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., Hilton Head Island, SC, USA, vol. 1, pp. 316–323 (2000)
19. Derraz, F., Taleb-Ahmed, A., Pinti, A., Chikh, A., Berekci-Reguig, F.: A geometrical active contour based on statistical shape prior model. In: Proc. 15th IEEE Int. Symposium Signal Processing and Information Technology, Sarajevo, Bosnia & Herzegovina, vol. 1, pp. 432–436 (2008)
20. Cremers, D., Osher, S.J., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *Int. J. Comput. Vis.* 69(3), 335–351 (2006)
21. Hossam, E., Munim, A.E., Farag, A.A.: A shape-based segmentation approach: an improved technique using level sets. In: Proc. 15th IEEE Int. Conf. Comput. Vis., San Diego, CA, USA, vol. 1, pp. 930–935 (2005)
22. Chen, Y.-T., Tseng, D.-C.: Medical image segmentation based on Bayesian level set method. In: Gao, X., Müller, H., Loomes, M.J., Comley, R., Luo, S. (eds.) MIMI 2007. LNCS, vol. 4987, pp. 25–34. Springer, Heidelberg (2008)
23. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. *Int. J. Comput. Vis.* 22(1), 61–79 (1997)
24. He, X., Niyogi, P.: The LPP algorithm. In: Advances in Neural Information Processing Systems 16 (NIPS), Vancouver, Canada, vol. 1, pp. 153–160 (2003)

An Improved Texture Feature Extraction Method for Tyre Tread Patterns

Ying Liu, Zong Li, and Zi-Ming Gao

School of Communication and Information Engineering,
Xi'an University of Posts and Telecommunications, Xi'an 710061, China
xupt2011@126.com

Abstract. Tamura features have been found to be effective in describing image textures and contrast is one of the Tamura features popularly used. As a global variable, contrast can well describe the statistical distribution of the brightness in the entire image, but cannot reflect the local brightness information of the image. To solve this problem, this paper proposes an improved texture feature extraction method which makes use of the statistical moments of intensity histogram to extract more information from the image. Tested on a tyre tread pattern dataset, the proposed method is found to be able to provide better retrieval performance than other existing methods.

Keywords: Image Retrieval, Tamura feature, Contrast, Statistical moments.

1 Introduction

With the rapid popularization of the Internet and multimedia technology, the size of image collection is increasing rapidly. In order for intelligent and efficient image database management, content-based image retrieval (CBIR) have been developed since 1980's. CBIR performs image retrieval based on the similarity measure of image features such as color feature, texture feature, shape feature and spatial location. These features are extracted from image content[1].

Texture features describes the homogeneity of image surface and the spatial distribution of different elements not depending on the color and brightness information. It reflects the global and local structure of images and has been widely used for image retrieval. There are many texture feature extraction methods designed for CBIR, including wavelet-based texture feature, Tamura feature, feature Gray-Level Co-occurrence Matrix and so on[1]. Tamura texture features have six dimensions including coarseness, contrast, directionality, line-likeness, regularity and roughness, which are designed based on human visual system. The first three have been commonly used for image retrieval [1,2].

To further improve the performance of Tamura features for image retrieval, the author in [3] modified the definitions of coarseness and directionality. This work intends to find effective texture extraction method for type tread patterns as a special type of data for our project in public security area. This paper proposes a new method to calculate the contrast feature by making use of the statistical moments of intensity histogram to extract more information from the image. Intensive experiments on a tyre

tread pattern dataset were carried out to test the performance of the proposed method by comparing its retrieval performance with other existing texture feature extraction methods. The simulation results prove the effectiveness of the proposed method for tyre tread patterns.

The rest of this paper is organized as follows. Section 2 reviews Tamura texture feature and Section 3 describes the method we proposed. Section 4 presents the experimental results on tyre tread texture data set. Section 5 concludes this paper and suggest future research directions.

2 Review of Texture Feature Extraction

According to human visual perception, Tamura et.al. proposed six texture features in 1978, including: coarseness, contrast, directionality, line-likeness, regularity, roughness. The first three components are more related with human perception than the other three. In addition, the last three components can be derived from the first three. Hence, the first three Tamura texture features are more effective in describing image textures and are more often used for image retrieval[4-6].

2.1 Coarseness

Coarseness, as the basic feature of the Tamura texture, reflects the largest size of the texture elements, even where a smaller texture exists [3,7]. The larger the coarseness value is, the rougher the texture is. For the different texture pattern structures, the more primitive size or the less primitive number of repetitions, gives people the impression of roughness. The essence of calculating the value of coarseness is the sliding value of the pixels with different size windows. It can be summarized as follows:

Step1: Taking the average at every pixel over neighborhoods whose sizes are the power of 2. The average over the neighborhood of size $2^k \cdot 2^k$ ($k=0,1,2,3,4,5$) at every pixel is:

$$A_k(x, y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} g(i, j) / 2^{2k} \quad (1)$$

where $g(i, j)$ is the gray-level at (i, j) .

Step2: For each pixel, calculating the differences between the not overlapping neighborhoods on horizontal and vertical directions:

$$\begin{aligned} E_{k,h}(x, y) &= |A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y)| \\ E_{k,v}(x, y) &= |A_k(x, y + 2^{k-1}) - A_k(x, y - 2^{k-1})| \end{aligned} \quad (2)$$

Step3: For each pixel, selecting the best size which gives the highest output value:

$$S_{best}(x, y) = 2^k + 1 \quad (3)$$

where k maximizes E in either direction:

$$E_k(x, y) = \max(E_{k,h}(x, y), E_{k,v}(x, y)) \quad (4)$$

Step4: Taking the average of S_{best} over the picture to be the coarseness measure F_{crs} :

$$F_{crs} = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n S_{best}(i, j) \quad (5)$$

where m and n are the effective size of the image.

Apparently, F_{crs} describes the rough grain size characteristic of the texture image, but when textons of different sizes exist in the image, coarseness as defined in (5) can't well reflect the texture characteristic of the image as there is information loss. An improved method was proposed in [3] which describes the distribution of S_{best} using the histogram of coarseness, not just taking the average of coarseness. The histogram of coarseness conveys the information of textons of different sizes in the image and thus reflects the texture features of different regions. Thus, it brings performance improvement in the image retrieval.

2.2 Directionality

Direction[3,8] is the most basic feature of the image which contains a large amount of image information. It describes globally how the texture in the image is distributed or concentrated along certain orientations. The calculation of directionality can be summarized as following:

Step1: Calculating the gradient vector at each pixel which includes its modulus $|\Delta G|$ and the edge directionality θ as:

$$\begin{aligned} |\Delta G| &= (\Delta_H + \Delta_V)/2 \\ \theta &= \tan^{-1}(\Delta_V/\Delta_H) + \pi/2 \end{aligned} \quad (6)$$

where Δ_H and Δ_V are the horizontal and vertical elements, calculated as the convolution of the image with the following 3· 3 operators :

$$\begin{array}{c} \begin{vmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{vmatrix} \quad \begin{vmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{vmatrix} \end{array} \quad (7)$$

Step2: Counting all pixels with $|\Delta G| \geq t$ and quantizing θ by

$(2k-1)\pi/2n \leq \theta \leq (2k+1)\pi/2n$, we obtain the number of the points $N_\theta(k)$ which satisfy the above constraints. Then, building the edge probabilities histogram HD :

$$HD(k) = N_\theta(k) / \sum_{i=0}^{n-1} N_\theta(i), \quad k = 0, 1, \dots, n-1 \quad (8)$$

In our experiments, we used $n=36$ and $t=12$. The histogram for obvious directional image will exhibit a peak, for images without obvious direction is relatively flat.

Step3: Finally, the overall direction can be obtained by calculating the histogram peak sharpness, This measure can be defined as follows:

$$F_{dir} = \sum_p^{n_p} \sum_{\phi \in w_p} (\phi - \phi_p)^2 H_D(\phi) \quad (9)$$

where p is the peak value of histogram, n_p are all the peak values of the histogram. For each p , w_p represents all the bins which include it, and ϕ_p is the bin which has the highest peak value.

Directionality as a global property over the given region, measures the total degree of directionality, but cannot effectively reflect the local different orientations or patterns. As in [3], the author uses Fourier Transform to convert H_D to frequency domain. Then choose the modulus instead of F_{dir} to describe the texture. The advantage of this method is that it can eliminate the phase difference between the image histograms and thus the texture feature is rotation-invariant.

2.3 Contrast

Contrast reflects the statistical distribution of brightness of pixels in the image and is determined by four factors: gray-level dynamic range, polarization degree of the white and black part in the histogram, sharpness of the edge, repeat model cycles [3,13]. Often we define contrast as following,

$$F_{con} = \sigma / (a_4)^n \quad (10)$$

where $a_4 = \alpha_4 / \delta^2$, α_4 is the fourth moments, δ^2 is the variance. Experimentally, $n=1/4$ was the best value obtained by Tamura et.al.

3 Description of the Proposed Method

As described above, the authors in [3] designed new formulas to improve the Coarseness and Directionality of Tamura features. Here, we present a new method to calculate Contrast.

From (10), we can see contrast is a global variable which can well describe the statistical distribution of the brightness in the entire image, but cannot reflect the local brightness information of the image. We propose an improved method, which makes use of the statistical moments of intensity histogram to extract more information from the image. The six different features describing the intensity distribution in the image are as follows, with the expression for the n th moment about the mean given by:

$$\alpha_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i) \quad (11)$$

Then, we can obtain the six different texture feature descriptors as:

Mean: A measure of average intensity, which represents the brightness information of image.

$$m = \sum_{i=0}^{L-1} z_i p(z_i) \quad (12)$$

where z_i is a random variable intensity, $p(z_i)$ is the histogram of the intensity levels, L is the number of possible intensity levels.

Standard deviation: Measure the contrast of gray level intensities as:

$$\delta = \sqrt{\infty_2(z)} = \sqrt{\delta^2} \quad (13)$$

where δ^2 is the variance, and is the second moments $u_2(z)$.

Smoothness: Using standard deviation value, measures the relative smoothness of the intensity.

$$R = 1 - 1/(1 + \delta^2) \quad (14)$$

Third moment: Measures the skewness of a histogram.

$$\infty_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i) \quad (15)$$

Uniformity: Measures the distribution of intensity level.

$$U = \sum_{i=0}^{L-1} p^2(z) \quad (16)$$

Entropy: Measures the randomness pixels value of the distribution.

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \quad (17)$$

The three images in Fig.1 (from left to right) are examples of images of smooth, coarse and periodic tire tread patterns. The histogram of these images are shown in Fig.2. Then use (12) to (17) obtain six different values to construct a new texture feature: $T = [m, \delta, R, \infty_3, U, e]$ for the better result.



Fig. 1. Examples of different tire tread patterns

Apparently, the results are in general agreement with the texture content of the respective images. Above all, through these texture feature descriptors to extract tire tread pattern image texture characteristics have obvious difference, thus these feature descriptors are used to describe automobile tire texture feature is effective and feasible.

Table 1. Texture features for the images shown in Fig.1

TEXTURE	SMOOTH	COARSE	PERIODIC
Mean	63.0027	120.5328	97.8410
Standard Deviation	14.8022	46.5564	34.8111
Smoothness	0.0034	0.0323	0.0183
Third Moment	-0.0211	-0.6853	-0.0976
Uniformity	0.0202	0.0068	0.0084
Entropy	5.8866	7.3744	7.0188

4 Experimental Results

To verify the performance of the proposed algorithm in image retrieval, a set of experiments are carried out on a tyre tread pattern dataset from a Public Security Criminal Investigation database. The dataset contains 200 images in 40 categories with 5 similar images in each category. Inner query is used to test retrieval performance, that is, images from the dataset are selected as queries. Texture features from all the images are extracted and are normalized. In our experiments, Euclidean distance has been proved to be more effective with higher retrieval accuracy and better ranking order. Compared with other similarity measure methods such as Canberra distance used in[3], Image similarity measure [10] is defined as the Euclidean distance between texture feature vectors as below:

$$d_c = \sqrt{\sum_{i=1}^L (t_q(i) - t_d^j(i))^2} \quad (18)$$

where d_c is the distance between feature vectors, $t_q(i)$ is the feature vector of the query image. $t_d^j(i)$ is feature vector of the database image, L is the length of feature vector[1].

In order to further improve the accuracy of retrieval, we integrate coarseness, directionality as in [3] and the proposed contrast together as texture feature of images , and the distance between textures feature vectors is calculated as:

$$d = (1/3) \cdot d_{coa} + (1/3) \cdot d_{dir} + (1/3) \cdot d_{con} \quad (19)$$

Precision is used to evaluate the performance of image retrieval as:

$$precision = s/k \quad (20)$$

where s is the number of relevant images retrieved, k is the total number of images returned.

In our first set of experiments, we compare the performances of the proposed method applied to image blocks of sizes 4*4, 8*8, 16*16, 32*32, 64*64, and the entire image as one block. The results in Fig. 2 show that dividing the image into small blocks and calculating the average texture feature does not bring any improvement in retrieval performance.

Then, we compare retrieval performance of three texture features: the proposed method, the method provided in [3], and wavelet-based texture features. In our experiments, we applied 2-level wavelet transform with Bior2.2 basis to obtain texture feature of 14 dimensions. Details of wavelet-based texture feature can be found in [2].

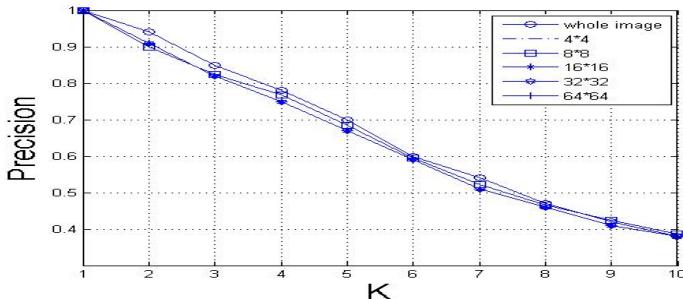


Fig. 2. Compare the top 10 images returned using different block sizes methods

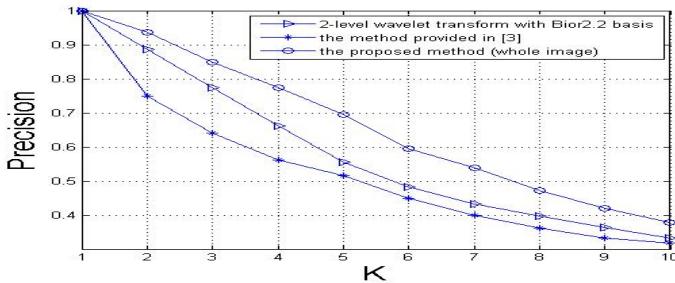


Fig. 3. Precision-K

Fig.3 compares the precision of the three methods. We can conclude that the proposed method provides the best retrieval performance. For example, when $k=5$, using our method, there is an obvious improvement in the retrieval precision from 57.8% to 70.1% compared with the method in [3]. Fig.4(b)-(d) displays the top 10 images returned using different texture features, given an example query as in Fig.4(a). It can be seen that the rank of the returned relevant images has been improved using the proposed method. For example, when $k=5$, using our method, the 5 returned images are all relevant to the query; using the method in [3], there are 3 relevant images; using wavelet-based texture features, there are only 2 relevant images among the first 5 returned.

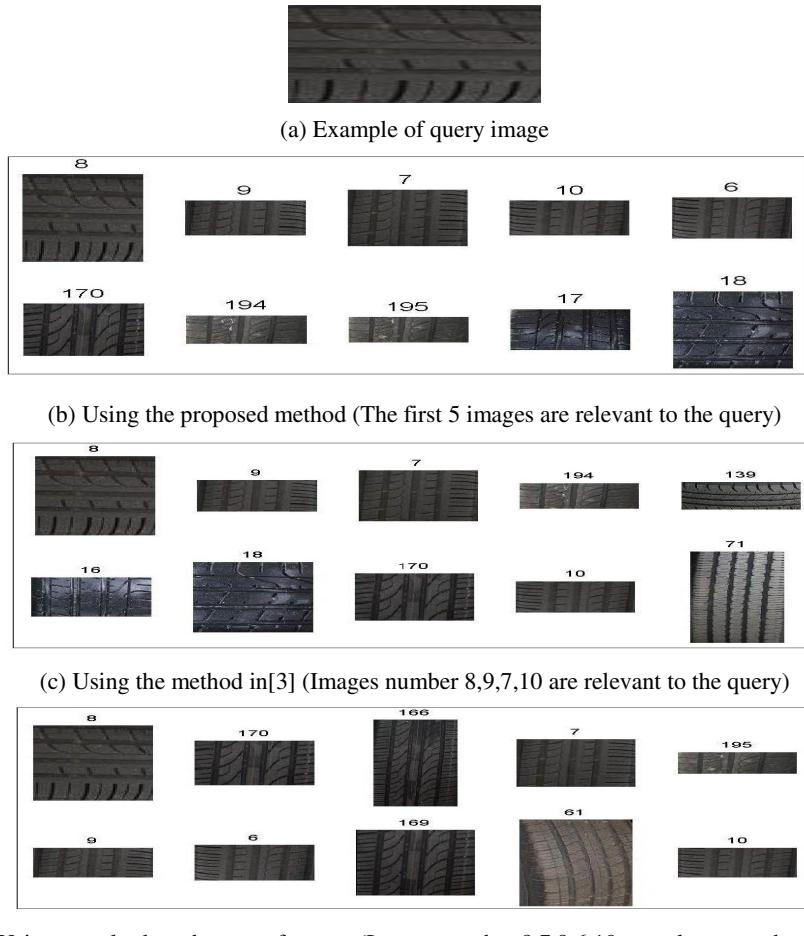


Fig. 4. Top 10 images returned using different methods

5 Conclusion and Future Work

This paper presented an improved texture feature extraction method for tyre tread patterns which includes a new definition of Contrast. This method made use of the statistical moments based on intensity histogram to extract more information from the image. Experimental results proved that the proposed method not only provides promising retrieval performance but also is efficient in computation. In our future work, we will further optimize the proposed method for application in larger dataset, by introducing more advanced methods such as machine learning.

Acknowledgements. This work was supported by National Natural Science Foundation of China project No. 61202183 Shaanxi Province Education Office project No.12JK0504 , as well as Shaanxi ‘100 Experts’ Plan.

References

1. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high level semantics. *Pattern Recognition* 40, 262–282 (2008)
2. Liu, J.: Technology of texture feature extraction based on wavelet. *Computer Engineering and Design* 13, 3141–3144 (2007)
3. Hao, Y., Wang, R., Ma, J., Zheng, J.: Image retrieval based on improved Tamura texture features. *Science of Surveying and Mapping* 4, 136–138 (2010)
4. Tamura, H., Mori, S., Yamaeaki, J.: Texture features corresponding to visual perception. *IEEE Trans. on Systems, Man and Cybernetics* 6, 460–473 (1978)
5. Wang, S., Qi, C., Cheng, Y.: Application of Tamura texture feature to classify underwater targets. *Applied Acoustics* 2, 135–139 (2012)
6. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: *Digital Image Processing Using MATLAB*, pp. 285–363. Publishing House of Electronics Industry, Beijing (2009)
7. Schaefer, G.: Content-based image retrieval: Some basics. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) *Man-Machine Interactions 2*. AISC, vol. 103, pp. 21–29. Springer, Heidelberg (2011)
8. Zhu, Z.-L., Zhao, C.-X., Hou, Y.-K., Fan, Y.: Rotation-invariant texture image retrieval based on multi-feature. *Journal of Nanjing University of Science and Technology* 36, 375–380 (2012)
9. Liu, Z., Wada, S.: Robust feature extraction technique for texture image retrieval. In: *Proceedings - International Conference on Image Processing, ICIP*, Genova, Italy, pp. 525–528 (2005)
10. Majumdar, I., Chatterji, B.N., Kar, A.: Texture feature matching methods for content based image retrieval. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)* 24, 257–269 (2007)

Multi Gesture Recognition: A Tracking Learning Detection Approach

Meng-Yuan Shi and De-Chuan Zhan

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{shimy,zhandc}@lamda.nju.edu.cn

Abstract. Many real world activities involve the interactions of multiple gestures, e.g., jogging on the playground with both legs and waving hands, saying good bye with shaking both hands, etc. However, current vision based gesture recognition algorithms assume there is only single gesture in the scenario. Some existing multiple gesture recognition detection systems require the aid of particular devices such as multi-touch pad, infrared sensors, gyroscope sensors etc. In this paper, we proposed a new tracking learning detection framework for recognizing the multiple gestures in the video stream. The framework is based on tracking learning detection (TLD) [1] approach, which integrates the short-term gesture tracker and online learned gesture detector. With the collaboration of tracker, detector and online learning algorithm in TLD, it can be successfully adapted to vision based multi-gesture recognition. Experiments show that our framework outperforms detection based methods with vision based multi gesture recognition.

Keywords: Multi-gesture recognition, object tracking, object detection, machine learning.

1 Introduction

The gesture recognition in Human Computer Interaction (HCI) has become an attractive research fields in the past few years since there are increasingly real applications rely on the natural gestures. Human gestures, compounded with physical movements and postures of the fingers, hands, arms or bodies, are expressive for conveying meaningful information and forms a major part for information communication in our everyday life.

Most recent developed gesture recognition approaches focus on the single gesture. However, in the real case there are few occasions that only single gesture play the role in video streams, e.g., people jogging on the playground waving hands and legs simultaneously, waving both of the hands for a while expresses the meaning of saying goodbye etc. Previous work on gesture recognition, such as mathematical models based on Hidden Markov Model [2] and approaches based on Finite State Machine mainly focus on single gesture recognition tasks. As a consequence, these methods cannot handle the cases that different kinds of gestures appear in the video stream simultaneously. The simple case like a hand gesture recognizers for the action

“stretch” and “grab”, i.e., only one kind of gesture would appear in the video at one time can be recognized by these approaches. If people grab on the left hand and stretch on the right hand to express the meaning he wants to grab the objects on the left hand and put down the objects in the right hand. Single gesture recognizer is insufficient for recognizing this.

The main challenge that multi-gesture recognition brings is the trajectory of the multi-gestures should be correctly recorded and recognized, since in the single gesture recognition task, the gesture can be determined by only exploiting the information of the initial status and the final status, while for multi-gesture recognition tasks, it is required to make a distinction between the final status of different gestures, i.e. to recognize action “saying goodbye” the trajectories and the appearance of both hands gestures should be recorded and evaluated, the correspondences between the kinds of initial status and the final status should be provided. Recently, with the emergence of special designed devices like multi-touch pad, gyroscope and infrared sensors, people can easily locate the position of the hands and catch the trajectory of multiple finger-prints. Furthermore, these devices make detection of human gesture feasible task. However, these particular devices based multi-gesture recognition cannot provide a natural human computer interaction environment for the end users. Since the gesture recognition methods are strongly depended on the certain devices, consequently the end user has to wear the sensors or work on the multi-touch pad.

In this paper, possibilities of vision based multi-gesture recognition are dis-cussed. We emphasize in multi-gesture recognition, the trajectories for different ges-tures should be tracked; the status of different gestures should be detected as well. Therefore a tracking learning detection multi-gesture recognition framework is proposed.

The rest of the paper is organized as follows. We start by giving a brief review of the related work on gesture recognition. Then we formulated multi-gesture problem and introduce TLD framework. Finally, we compare the TLD method with some detection based methods and gives the conclusion

2 Related Works

Researchers have devoted to the gesture recognition for many years and a comprehensive review of gesture recognition methods can be found in [3]. Here we only reviews the tracking, learning and detection techniques used in the gesture recognition.

Tracking approaches provides the trajectory features for recognition tasks. Earlier gesture recognition use common motion features such as optical flows [4] and motion trajectories [5] for gesture representation. Recently various spatio-temporal features and descriptors are proposed [6,7]. Laptev adopt Histogram of Oriented (HOG) like features to video feature extraction[6], and propose a method to detect Space-Time Interest Points (STIP) and use Histogram of Oriented Optical Flows (HOF) as descriptors, and it achieves the state-of-the-art performance on action recognition. In[7], Dollár et extract the descriptors from space-time cuboids from temporal Gabor-filters. Chaudhryet al.[4] calculates a sequence of HOF and use Binet-Cauchy kernels on nonlinear dynamical systems.

The learning techniques are employed to enhance the performance of gesture classifier. The mainly used learning models are time-series based models. e.g., the Hidden Markov Model (HMM) has been widely used for gesture recognition. Marcel et al[2] used EM to train an input-output HMM, and apply it to recognize gestures from the contours of the hand, which was extracted by image segmentation and hand tracking. Delgado[8] mixed HMM and Artificial neural network(ANN) to identify the trajectories of different gestures. In [9], vector quantization were adapted to transform feature vectors to symbolic sequences, and subsequently modeled by HMM. The Finite State Machine (FSM) is another well received model for gesture recognition[12]. With fingertips detected to extract feature vectors, Davis[10] apply FSM to model the four phases of a gesture. In [12], the position of detected head and hands were used as training feature for determining the states of FSM. Besides, the topologies of a self-organizing-mapping neural network were used by Flórez[11] to determine hand postures and gestures.

Template matching was the first method employed for detecting hands, however it was not invariant to scale and rotation. To cope with the variability due to scale and rotation, some authors have proposed scale and rotational normalization methods (e.g.[13]). Several appearance based methods[14,15] attempt to detect hands gestures based on hand appearances.

3 Tracking Learning Detection for Multi Gesture Recognition

3.1 Basic Idea

Different to single gesture recognition, multi gesture recognition calls for the collaboration of different gestures. Therefore, in this paper, we emphasize that the features of multi-gesture recognition should be related to the trajectories of multiple gestures. In order to record the full trajectories for each gesture, the tracking, learning and detection should be employed simultaneously.

Tracking estimates the motion of gestures between consecutive frames under the assumption that the gesture is visible in the video stream and frame-to-frame variance is limited. Such assumption seldom occurs in real-world multi-gesture recognition. Since the objects would often disappear in the video stream, and the short term tracker seldom considers the dynamic change of appearance of gesture. However, the result of tracking gives a prediction of the possible positions for gesture, and they can be the candidates of the gesture positions.

Besides, in order to obtain the status of gestures, the possible regions that contain the target gestures are necessarily captured at first. We utilize detectors to take efforts on each frame and perform full scanning of the video frames. The advantages of making use of detectors are that detectors are not limited to the assumption that frame-to-frame gesture variance is limited, and it is immune to disappearance of gestures in the video stream. However, the performance of the detector largely depends on the training set and cannot catch the dynamic change of gestures in the video stream.

Learning observes performance of both tracker and detector, estimates the errors of them and produces training examples to refine the tracker and detector. The refined

tracker and detector are able to avoid such detecting and tracking failures, and have a better understanding of the gestures. With the collaboration of learning, tracker and detector, the final gesture detector generalizes to a subspace fully describes gesture.

So, we make use of TLD (tracking learning detection) methods and propose a new framework for extracting the trajectories of different gestures. TLD was first proposed in[1] by Kalal, and has achieved excellent performance in long-term object tracking and face recognition. In this paper, we claim it also an excellent framework for multi-gesture recognition. The architecture of the framework could be found in Fig.1.

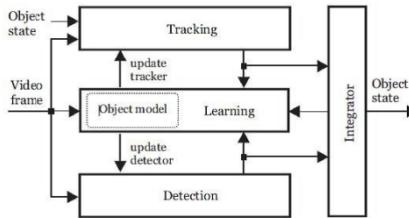


Fig. 1. Detailed block diagram of the tracking learning detection (TLD) framework

3.2 Tracking Learning Detection (TLD) Framework

Lucas-Kanade (LK) method[13] is used as short term gesture tracker to implement the TLD framework. The sampled feature points in the bounded box are tracked. The trajectory are recorded for further tracking. The recorded trajectory will be analyzed by two kinds of events, “growing event” and “pruning event” for further tracking. The “growing events”, the events identified as examples likely to be the target gesture, provides abundant views of the training set. While the “pruning events”, which are considered misclassified by the online tracker and detector, are regarded as negative examples for the training set, purify the model with “negative feedback”. The dynamic interaction of “growing events” and “pruning events” makes the system stable, and ensures the L_t converge to the subspace L^* describing the gesture. For the gesture detector, we adopt sequential random forest algorithm and 2BitBP feature. Sequential random forest enables an online learning approach for embedding the growing and pruning events, the 2BitBP feature catches the texture information of image patch and can be fast extracted.

3.3 TLD for Multi-gesture Recognition

Let $G = \{g_1, g_2, \dots, g_n\}$ denote the gestures that need to be recognized, where n is the number of the gestures for recognition. The video sequence is represented as an image sequences $V = \{I_1, I_2, \dots, I_m\}$, where m is the number of frames in the video.

The state of gesture i at time t is denoted as 3-tuples $s_{it} = (x_{it}, y_{it}, a_{it})$, x_{it}, y_{it} represents the x-coordinate and y-coordinate of the gesture i at time t . a_{it} tells whether the gesture appears or not 1 for the gesture appear, 0 for not, when the gesture doesn't appear in the video stream the x_{it}, y_{it} would be 0. Each gesture g_i is described as a vector $g_i = (s_{i1}, s_{i2}, \dots, s_{im})$, where i denotes the gesture Id, m is the length of the video stream. s_{it} is the 3-tuple representing the state of the gesture i at

time t . Finally, the whole video is described by a $m*n$ vector $v = (g_1, g_2 \dots g_n)$, each g_i is a m-length vector describing the states of the i_{th} gesture.

Different from previous reviewed descriptors which only focus on the initial and final state of the gestures. In multi-gesture system, our proposed features embedding the trajectories of multi-gestures, so the performance of the gesture tracker plays a key role in multi-gesture recognition. If the gesture tracker fails, there would be little difference among different gesture features. Such features will affect the classification result of the gesture recognizer. In this paper, TLD is employed as tracker for extracting feature representing the video.

With generated features using TLD-base gesture tracker. The final gesture recognition problem could be formulated as a supervised learning problem. The training set is a set of frames containing multiple gestures and is denoted as a video instance, $V = \{(v_1, y_1), (v_2, y_2) \dots (v_n, y_n)\}$, and gallery gestures $\{g_1, g_2 \dots g_m\}$, where v_i denotes the video instance containing multi-gestures. g_i is the gestures set that the video contains. y_i is label for the video. Given dataset $D = \{(v_1, y_1), (v_2, y_2) \dots (v_n, y_n)\}$, a learning algorithm seeks an function $h: V \rightarrow Y$ to predict the label of new coming instance v . Since there are multiple labels for the video clips, we use multi-class classifier as the final classifier.

Algorithm 2. TLD for gesture recognition

- Input:** video sets $V = \{(v_1, y_1), (v_2, y_2) \dots (v_n, y_n)\}$, gallery gestures $\{g_1, g_2 \dots g_m\}$
1. Use TLD gesture tracker to record the trajectory for each gestures, in the videos and construct gesture descriptor.
 2. Concatenate the descriptor for different gestures, construct the video descriptor
 3. Train multi-class classifier by labeled videos and video descriptor.
-

4 Experiments

Previous works on gesture recognition don't concern the problem of multi-gesture recognition, and apparatus based multi-gesture recognition problem is different from our methods. To testify the performance of proposed TLD framework, we compared our method with other detection based algorithms. All the methods are used for generating features describing the multi-gesture semantic. The classification accuracy on these features is compared.

The dataset contains 10 categories of dynamic multi-gesture semantics in total i.e. *grab*, *put down*, *lift up*, *victory*, *give up*, *clockwise circle*, *chest extension*, *showing muscle*. Each example is represented using 25 images depicting the dynamic motion of the gesture. Fig 2 shows the semantics of give up, lift up and turn over. We could see that all these gestures calls for the interaction of 4-gestures: the left hand stretched left hand with fist, right hand stretched and right hand with fist. The back-ground is complex and diverse. The condition is common in real world multi-gesture applications. Each semantic contains 30 samples to our dataset.



Fig. 2. Samples from our dataset

Table 1. Accuracy of different classifiers on different extracted features

Final Classifier	cascade	SVM	Random-forest	TLD
Random-forest	55%	52%	58%	88%
Adaboosting	58%	58%	55%	86%
SVM	56%	49%	56%	85%



Fig. 3. Tracking of left/right hand stretched in turn over

In experiments, three detection based algorithms cascade[14] gesture detector, SVM based detector[16] and random forest based detector[17] are compared with TLD. For each compared method, 4 detectors are trained to detect 4 kinds of gestures (left hand stretched, left hand fist, right hand fist and right hand stretched). The training data for each detector contains 90 positive samples and 2000 negative samples. The HOG descriptor extracted from each sample is used as input for random forest and SVM based detector.

After gesture detector extracts the trajectories for each gesture. We use methods described in section 3.3 for constructing the features for final multi-gesture classification. The final feature encodes the trajectory information of each gesture. If the gesture detector performs with low accuracy or the trajectory couldn't be detected by the gesture detector, there would be little difference among features for different semantics. So the detection and tracking performance of the detectors plays a key role in multi-gesture recognition. To eliminate the effect of final classifier, three classic machine learning algorithms (adaboost, SVM, random forest) are adopted as final classifier for multi-gesture classification. Base classifier in adaboost is decision-stump.

In the experiment, the average accuracies of classifiers for each generated feature are evaluated using 10 times cross-validation (CV). The final classification result is listed in Table 1, the background is excluded during the process of multi-gesture recognition. The detection of left/right hand stretched in semantic turnover is shown in figure 2. In the semantic, the background is excluded during the process of gesture detection. From the table, it can be seen that TLD performs significantly better than all the other detection based algorithm. Besides TLD doesn't need any offline training process, one picture containing the gestures to be tracked is sufficient for the initialization of TLD. Since detection based algorithms suffers from tracking or detection failure, and the detection based algorithm is unable to catch the dynamic changes in the appearance of the gestures in the video stream. However such case is common in multi-gesture recognition. Since we adopt three kinds of classifier for final multi-gesture classification. The results of Table 1 also prove that the result is invariant to the selection of final classifier. The Fig 3 shows the tracking result of TLD with gesture left/right hand stretched under semantic turn over. From the figure, we could see the TLD is able to locate the positions of different gestures, although the gesture appearance will be changed during tracking.

5 Conclusion

Nowadays, large amounts of gesture recognition applications involving the interaction of multiple gestures. However, previous work on gesture recognition algorithms assumes there is only one gesture in the video data, and the recognition of multi-gesture often calls for the usage of apparatus such as multi-touch pad, infrared sensor or gyroscope sensors. However, the apparatus based multi-gesture recognition algorithms could not provide a natural human computer interface for the end user. Complete vision based multi-gesture hasn't been explored in the past decades. In this paper, we adopt TLD framework which has been successfully adapted to long-term object tracking to multi-gesture recognition. In the framework, tracking is used for short-term tracking of gestures. Online learning with "growing events" and "pruning events" iteratively elaborated the detector model and enables it to fully represent the gesture appearance. Gesture detector runs in parallel with tracking and enables re-initialization of gesture tracker in case of tracking failure. With the combination of tracker, learner and detector the system converges to a subspace describing the gesture. The trajectories of different gestures are extracted as features describing the video stream, and the final classifier is trained on it for future classification. We compared the TLD framework with other detection based algorithms on one multi-gesture artificial dataset, experiments validates the performance of TLD framework on multi-gesture recognition. In the future work, we will employ more algorithms for different parts of TLD framework, and redefine useful "growing events" and "pruning events" for multi-gesture recognition. Besides, current work only focus on static gestures, we will try to extend it for dynamic gestures in the future.

Acknowledgments. This research was supported by NSFC (61105043) and JiangsuSF (BK2011566).

References

1. Kalal, Z., Matas, J., Mikolajczyk, K.: Online learning of robust object detectors during unstable tracking. In: IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops (2009)
2. May, P., Ehrlich, H.-C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow Through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Mitra, S., Acharya, T.: Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37(3), 311–324 (2007)
4. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1932–1939 (2009)
5. Yang, M.H., Ahuja, N., Tabb, M.: Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(8), 1061–1074 (2002)
6. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 432–439 (2003)
7. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (October 2005)
8. Rajko, S., Qian, G., Ingalls, T., James, J.: Real-time gesture recognition with minimal training requirements and on-line learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (June 2007)
9. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden Markov model. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1992, pp. 379–385 (June 1992)
10. Davis, J., Shah, M.: Recognizing hand gestures. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 800, pp. 331–340. Springer, Heidelberg (1994)
11. Flórez, F., García, J.M., García, J., Hernández, A.: Hand gesture recognition following the dynamics of a topology-preserving network. In: Proceedings of 5th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 318–323 (May 2002)
12. Freeman, W.T., Weissman, C.: Television control by hand gestures. In: Proc. of Intl. Workshop on Automatic Face and Gesture Recognition, pp. 179–183 (June 1995)
13. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision* 56(3), 221–255 (2004)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511. IEEE (2001)
15. Cui, Y., Weng, J.J.: Hand sign recognition from intensity image sequences with complex backgrounds. In: Proceedings of IEEE the Second International Conference on Automatic Face and Gesture Recognition, pp. 259–264 (October 1996)
16. Liu, Y., Gan, Z., Sun, Y.: Static hand gesture recognition and its application based on support vector machines. In: IEEE 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD 2008, pp. 517–521 (August 2008)
17. Liu, Z., Xiong, H.: Object detection and localization using random forest. In: IEEE Second International Conference on Intelligent System Design and Engineering Application (ISDEA), pp. 1074–1078 (January 2012)

Kernel-Based Representation Policy Iteration with Applications to Optimal Path Tracking of Wheeled Mobile Robots

Zhenhua Huang, Xin Xu, Lei Ye, and Lei Zuo

College of Mechatronics and Automation, National University of Defense Technology,
Changsha, 410073, P.R. China
zhenhuahuang10@gmail.com, xinxu@nudt.edu.cn

Abstract. How to improve the generalization and approximation ability in reinforcement learning (RL) is still an open issue in recent years. Aiming at this problem, this paper presents a novel kernel-based representation policy iteration (KRPI) method for reinforcement learning in optimal path tracking of mobile robots. In the proposed method, the kernel trick is employed to map the original state space into a high-dimensional feature space and the Laplacian operator in the feature space is obtained by minimizing an objective function of optimal embedding. In the experiments, the KRPI-based PD controller was applied to the optimal path tracking problem of a wheeled mobile robot. It is demonstrated that the proposed method can obtain better near-optimal control policies than previous approaches.

Keywords: Reinforcement learning, path tracking, kernel methods, wheeled mobile robots, approximation policy iteration.

1 Introduction

Reinforcement learning (RL) [1] is one of the most important branches in machine learning. In RL, it is concerned with how an agent improves decisions at each step to maximize its long-term total reward, by interacting with the environment. To deal with the impact of model uncertainties and unknown disturbances, RL has been applied in the feedback control of mobile robots and drew more and more attention of researchers in robotics and engineering [2,3]. However, RL algorithms have the curse of dimensionality, the exponential growth of the number of parameters to be learned with the size of any compact encoding of system states [4]. To improve the generalization ability of RL algorithms requires the study of approximate RL theories and algorithms based on approximate value functions or policies.

One major technique for approximate RL is called value function approximation (VFA) [5]. According to the architectures of function approximators, there are two different kinds of VFA methods: linear [6] and nonlinear [7]. Although RL with nonlinear VFA exhibits better approximation ability than linear VFA, the

empirical results of RL applications using nonlinear VFA commonly lack a rigorous theoretical analysis and the nonlinear features are usually determined by manual selection. Linear approximation architectures, in particular, have been widely used as they offer many advantages in the context of value-function approximation. One common drawback of previous work in VFA is that the basis functions or kernel functions are usually hand-coded by human experts, but not automatically constructed from the geometry of the underlying state space.

Recently, based on the principle of Laplacian eigenmaps, a framework for VFA called proto-value function (PVF) was proposed [8], where the representation policy iteration (RPI) algorithm was developed. PVFs can be automatically constructed by using spectral analysis of the self-adjoint Laplacian operator. In MDPs with continuous or large state space, one challenge is how to choose a subset of samples from which a graph can be built [8]. In RPI, trajectory-based subsampling methods are used to select the subset from the collected samples. In order to reflect the topological structure of the whole state space better, the representative subset should be selected. However, the subsampling methods in RPI mostly aim at the subset's distributing uniformity other than its representation.

In this paper, a kernel-based RPI (KRPI) method, which has more efficient representations via kernel functions, is proposed. The intuition of the kernel trick is to map the data from the original input space to a higher dimensional Hilbert space as $\varphi : x \rightarrow F$. Then the representation using graph Laplacian is performed in this new feature space. The first novelty of KRPI is using the clustering-based subsampling method for the selection of the representative subset. Besides, it can be well applied to the algorithms that only need to compute the inner product of data pairs $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$. Then a kernelized graph Laplacian (KGL) is obtained in the higher dimensional Hilbert space F . The basis functions for VFA are calculated from the eigenfunctions of the KGL. Through the efficient representation in the kernel space, the approximation ability of the basis functions is improved and the control performance of the learned policy is also enhanced.

Over the past decades, wheeled mobile robots (WMRs) have received considerable attention in various industrial and service applications. These applications require mobile robots to have the ability of tracking specified paths stably and precisely. In this paper, a KRPI-based PD controller (KRPI-PD) is proposed to test the generalization and approximation ability of the KRPI method and then applied to the optimal path tracking of Wheeled Mobile Robots (WMRs).

The paper is organized as follows. In Section 2, a brief overview of MDP is introduced. Then, the KRPI approach for optimal path tracking is presented in Section 3. In Section 4, experimental results are provided to illustrate the effectiveness of the proposed method. The conclusion is drawn in Section 5.

2 Technical Background

The underlying formalism for many reinforcement learning algorithms is the Markov decision process (MDP) [1]. A Markov decision process is denoted as a

tuple $\{S, A, R, P\}$, where S is the state space, A is the action space, P is the state transition probability, and R is the reward function. The policy of the MDP is defined as a function $\pi: S \rightarrow P_r(A)$, where $P_r(A)$ is a probability distribution in the action space. The objective is to estimate the optimal policy π^* satisfying

$$J_{\pi^*} = \max_{\pi} J_{\pi} = \max_{\pi} E_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \quad (1)$$

where γ is the discount factor and r_t is the reward at time-step t , $E_{\pi}[\cdot]$ stands for the expectation with respect to the policy π and the state transition probabilities, and J_{π} is the expected total reward.

The value function $V^{\pi}(s)$ for policy π is a function that tells, for each state s , what the expected cumulative reward will be of executing the policy π . In the discounted setting, the value function $V^{\pi}(s)$ and the optimal state value function for the optimal policy are defined as follows:

$$V^*(s) = E_{\pi^*} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right] \quad (2)$$

According to the theory of dynamic programming [4], the optimal value function satisfies the following Bellman equation:

$$V^*(s) = \max_a [R(s, a) + \gamma E[V^*(s')]] \quad (3)$$

where $R(s, a)$ is the expected reward received after taking action a in state s .

Closely related to the value function is the so-called state-action value function, which is usually used to improve the control policy. The state-action value function, $Q^{\pi}(s, a)$ defined in (4), gives the expected cumulative reward of performing action a in state s following the policy π .

$$Q^{\pi}(s, a) = E^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right] \quad (4)$$

The state-action value function satisfies the following Bellman equation [1]:

$$Q^{\pi}(s_t, a_t) = E[r(s_t, a_t) + \gamma \sum_{a_{t+1} \in A} p^{\pi}(s_{t+1}, a_{t+1}) Q^{\pi}(s_{t+1}, a_{t+1})] \quad (5)$$

where the expectation $E[\cdot]$ is with respect to the state transition probability.

The optimal state-action value function can be obtained by $Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a)$, and then the optimal policy can be obtained by $\pi^*(s) = \arg \max_a Q^*(s, a)$.

3 KRPI for Optimal Path Tracking

3.1 Framework of KRPI

Fig. 1 shows the flow chart of the kernel-based representation policy iteration (KRPI) method. The KRPI method proposed in this paper has two distinctive

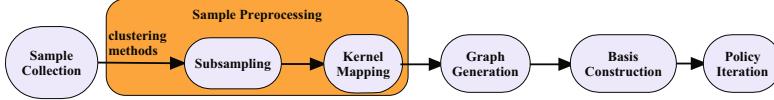


Fig. 1. The flow chart of KRPI

features when compared with representation policy iteration (RPI). One is the subset selection for the graph Laplacian construction. In KRPI, the selected data points using the clustering-based methods are much more representative than before. Besides, kernel tricks will be used in KRPI to map the original state space into the high-dimension feature space. The nonlinear kernel mapping is used to map the data into an implicit feature space F . Thus, a nonlinear subspace that can approximate the intrinsic geometric structure of the face manifold will be eventually gained.

Firstly, samples are collected randomly or by an initial policy. After sample collection, sample preprocessing which consists of subsampling and kernel mapping is one of the key elements in KRPI. Instead of the trajectory-based subsampling method proposed in [8], the clustering-based methods, such as K -means or Fuzzy C-means, are adopted for extracting the subsamples from the original collected samples. Then a nonlinear kernel mapping is used to map the state space of the subsamples into an implicit feature space F which may have very high or even infinite dimensions. Assuming a set of states $X = [x_1, x_2, \dots, x_n]$ in the subset D_x ; x_i is the collected i -th state of the agent. A nonlinear function φ to map the state datas into a high-dimensional feature space F : $\varphi(S) = [\varphi(s_1), \varphi(s_2), \dots, \varphi(s_n)]$.

After the nonlinear kernel mapping, a graph G can be constructed by connecting any two points in the subset and assigning the weights to the edges. Let $G = (V, E, W)$ denote an undirected graph with vertices V , edges E and weight matrix W whose entry w_{ij} means the weight on edge $(i, j) \in E$. E is the set of edges where $(i, j) \in E$ denotes an undirected edge from vertex s_i to vertex s_j . Samples obtained in the subsampling step construct the vertices of the graph and edges combine them by the weights. An undirected graph can be constructed by connecting two temporally successive states with a unit cost edge in discrete state space, or using a local distance measure such as k -nearest neighbor (knn) to connect states in continuous MDPs where weights are assigned to the edges as $w_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma)$, where $\sigma > 0$ is a predefined parameter, x_i and x_j are two connected state points in the constructed graph.

Let D be a diagonal matrix whose entries are the row sums of W . D can be seemed as the degree of vertices. The diagonal elements in $D = [d_{ii}]_{n \times n}$ can be represented by $d_{ii} = \sum_j w_{ij}$.

Then define $L = D - W$ as the graph Laplacian operator. From the constructed graph G , the learned basis functions $\{\phi(x_1), \phi(x_2), \dots, \phi(x_n)\}$ by RPI can be obtained by minimizing the following objective function as:

$$\phi^* = \arg \min_{\phi^T D \phi = 1} \phi^T L \phi \quad (6)$$

where D and W have been defined as above. Because the linear transformation $\phi = X^T \omega$, the objective function can be transformed as follows:

$$\omega^* = \arg \min_{\substack{\omega^T X D X^T \omega = 1}} \omega^T X L X^T \omega \quad (7)$$

The solution is provided by the matrix of the eigenvectors corresponding to the lowest eigenvalues of the generalized eigenvalue problem $L\phi = \lambda D\phi$.

In KRPI, a nonlinear function φ to map the state datas into a high-dimensional feature space F . There exists a coefficient vector $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$. Assume the mapping function is $\omega = \varphi\alpha$ and K is the kernel Gram matrix with $K_{ij} = k(x_i, x_j)$, then the objective function after kernel mapping can be represented as follows:

$$\alpha^* = \arg \min_{\substack{\alpha^T \varphi^T \varphi D \varphi^T \varphi \alpha = 1}} \alpha^T \varphi(X)^T \varphi(X) L \varphi(X)^T \varphi(X) \alpha \quad (8)$$

$$= \arg \min_{\substack{\alpha^T K D K^T \alpha = 1}} \alpha^T K L K \alpha \quad (9)$$

Thus, this minimization problem can be converted to a generalized eigenvalue problem with a constraint condition $\alpha^T K D K \alpha = 1$. The eigenvectors corresponding to the smallest eigenvalues are the solution:

$$K(D - W)K\alpha = \lambda KDK\alpha. \quad (10)$$

The graph laplacian in the kernel feature space F can be defined as below:

$$L_k = K(D - W)K \quad (11)$$

For continuous MDPs, a linear function approximation scheme is often employed to represent the action-value function $Q^\pi(s, a)$ for a policy π with a set of k basis functions $\phi(s, a)$:

$$\tilde{Q}^\pi(s, a) = \sum_{i=1}^k \phi_i(s, a) w_i. \quad (12)$$

where k is the number of the basis functions, and w_i is the i^{th} coefficient which can be determined by solving a fixed-point approximation $T_\pi Q^\pi \approx Q^\pi$, where T_π is the Bellman backup operator.

Commonly, a sampling technique is used to estimate the $w : \tilde{w} = A^{-1}b$. where the matrix A and the vector b can be estimated as

$$\begin{aligned} A_{t+1} &= A_t + \phi(s_t, a_t)(\phi(s_t, a_t) - \gamma \phi(s'_t, a'_t))^T, \\ b_{t+1} &= b_t + \phi(s_t, a_t)r_t. \end{aligned} \quad (13)$$

where (s_t, a_t, r_t) denotes the t^{th} sample point generated by the agent using a random or guided policy.

3.2 KRPI-Based Path Tracking of WMRs

In this paper, we studied the kinematic tracking control problem, with the goal of minimizing the tracking error between the real and desired path by controlling the velocity of the mobile robot. The mobile robot platform used in the paper is a P3-AT wheeled mobile robot system. RL, as an adaptive learning control method, can be used to optimize controller performance without much *a priori* knowledge about the model of WMRs. Motivated by this idea, we consider to apply RL approaches with PD feedback control to achieve optimal path tracking.

The control law in the PD controller is defined as follows :

$$\omega_d(t) = k_p(t)e(t) + k_d(t)\dot{e}(t) \quad (14)$$

where $e(t) = x_d(t) - x(t)$, $x_d(t)$ is the desired position and $x(t)$ the actual position, $k_p(t)$ and $k_d(t)$ are time-varying PD coefficients. RL approaches are used to optimize the control parameters, i.e. $k_p(t)$ and $k_d(t)$. As is well known, most RL algorithms are based on the model of MDPs. Thus, it needs to model the path tracking problem as an MDP first before using the algorithm proposed in the next section. We define the position and orientation of the mobile robot

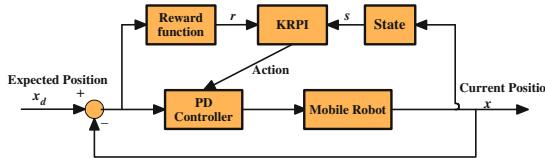


Fig. 2. PD control system based on KRPI algorithm

as the state space $S = \{(x_t, y_t, \theta_t)\}$. Our purpose is to optimize the PD control parameters $k_p(t)$ and $k_d(t)$, therefore we define the candidate parameters of the PD controller as the action space, i.e. $a \in \{(k_{p1}, k_{d1}), (k_{p2}, k_{d2}), \dots, (k_{pn}, k_{dn})\}$. According to the error between the actual path and the desired path, the reward function is defined as follows:

$$r(t) = \begin{cases} r_1, & |e(t)| \leq \varepsilon \\ -1, & |e(t)| > \varepsilon. \end{cases} \quad (15)$$

where $r_1 \geq 0$ is a constant reward and $\varepsilon \geq 0$ is an accepted boundary of the tracking error. Due to that the state transition probability P is unknown, we use API algorithms in this paper based on sampling. After learning, we can obtain an appropriate set of parameters for the PD controller when the robot is in a specific position and orientation. Fig. 2 illustrates the framework of the motion control system based on KRPI algorithm for WMRs.

4 Performance Evaluation

In this section, we perform experiments to demonstrate the effectiveness of the proposed KRPI method for optimal path tracking. In addition, we compare the

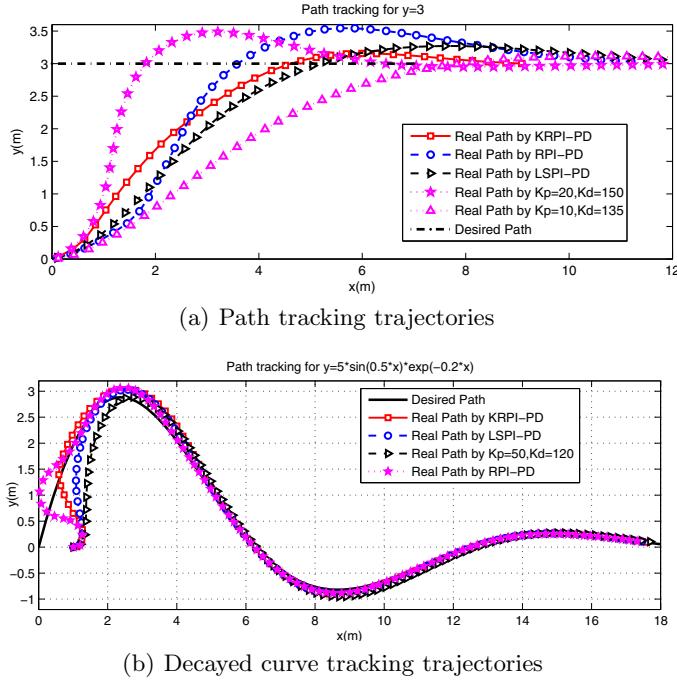


Fig. 3. Performance comparison in the tracking experiments

performance of KRPI with the different API algorithms, including LSPI [6] and RPI [8]. In all experiments, a global coordinate (x, y, θ) for the robot is defined, and the initial position of the robot is set in a small region around $(x, y)=(0,0)$. Because only lateral control is considered, the longitudinal velocity of the robot is set as a constant $v=0.2m/s$ in simulation and real time control. The action space of the MDP is defined as a series of PD coefficients expressed by $a \in \{(k_{p1}, k_{d1}), (k_{p2}, k_{d2}), (k_{p3}, k_{d3})\}$. The time step for online control is 0.5s.

The first reference path is a straight line: $y=3m$. Because of the difficulties in measuring the robot velocities, we only use the position and orientation information to define the state. The state vector of the MDP is defined as $s=(y, \theta)$, and the absorbing state is defined as $s_T \in \{(y, \theta) | |y - 3| < \beta_1, |\theta| < \beta_2\}$, where β_1 and β_2 are predefined parameters. The three actions for candidate PD parameters are defined as $a \in \{(20, 150), (5, 95), (10, 135)\}$. According to (14), the law in the PD controller is:

$$\begin{aligned} \omega_d(t) &= k_p(t)e(t) + k_d(t)\dot{e}(t) \\ &= k_p(t)(3 - y(t)) + k_d(t)(-v \sin \theta(t)). \end{aligned} \quad (16)$$

The learning control objective is to realize a near-optimal controller for the robot from the initial position to the absorbing state. To optimize the path-tracking performance, r_1 in the reward function 15 is -1.

In the experiment, β_1 and β_2 are set to be 0.05m and 0.05rad, respectively. To approximate the action value functions, a radial basis function (RBF) is used in LSPI. In the implementation of the three API algorithms, the discount factor is chosen as $\gamma=0.95$. The maximal number of the learning iterations is set 10. The number of the basis functions in KRPI and RPI are both set the same, such as 20. After the convergence of each algorithm, the performance of the final policy is tested on the real robot. The initial configuration of the robot is set as $(x,y,\theta)=(0,0,0)$. Fig. 3(a) illustrates the path tracking trajectories of different algorithms. For comparison, the best performance obtained by conventional PD control using two of the three candidate PD parameters is also shown.

The result indicates that the near-optimal policy of KRPI causes the robot to reach the absorbing state in a much fewer steps (shorter time), which is better than the other API algorithms and conventional PD control. The maximum deviation (MD), the response time (from the start state to the absorbing state) and the total tracking error for the different algorithms in the experiments are listed in Table 1(a). The learning controller based on KRPI not only made the robot get to the goal state in the least time, but also received the smallest MD and total tracking error compared with other controllers.

Table 1. The performance of different control strategies in the experiments

(a) The straight line tracking			
	Maximum Deviation (m)	Response Time (s)	Total Errors (m)
$K_p=20, K_d=150$	0.464	35.3	31.25
$K_p=10, K_d=135$	0.145	36.1	30.31
<i>LSPI-PD</i>	0.345	41.6	28.23
<i>RPI-PD</i>	0.526	41.3	27.42
KRPI-PD	0.121	31.2	20.53

(b) The decayed curve tracking			
	Maximum Deviation (m)	Response Time (s)	Total Errors (m)
$K_p=50, K_d=120$	0.241	45.6	36.27
<i>RPI-PD</i>	0.646	35.2	37.44
<i>LSPI-PD</i>	0.223	39.3	31.23
KRPI-PD	0.141	33.2	26.53

The second reference path is a decayed sine curve $y=5\sin(0.5x)\exp(-0.2x)$. In this experiment, the mobile robot starts at the initial point $(x,y)=(0,0)$ in the horizontal orientation. In the task, the state vector of the MDP is defined as $s=(x,y,\theta)$, and the three actions are defined as $a \in \{(30,200),(50,120),(15,180)\}$.

The parameters used in the curve line tracking are set the same as the straight line tracking, except for 15 basis functions used in KRPI and RPI.

The path tracking trajectories of different algorithms are shown in Fig. 3(b). Table 1(b) shows the maximum deviation, the response time and the total tracking error for the different algorithms in the experiments. After learning, the PD controller based on KRPI has better tracking performance than the other algorithms: within the least time tracking the goal curve with the smallest deviation.

5 Conclusion

In this paper, a kernel-based representation policy iteration (KRPI) method is proposed. The performance of KRPI is tested and evaluated on the path tracking control problem of a wheeled mobile robot. The principal advantage of KRPI is that it uses kernel trick to map the state space into a high-dimensional feature space and the Laplacian operator in the kernel-induced feature space can be obtained. In the path tracking experiments on the P3-AT robot platform, the robot controlled by the adaptive PD strategy learned from KRPI can track the goal line or curve with better performance than previous methods. Future works may include the performance evaluation of the proposed approach in the navigation of WMRs in complex environments.

Acknowledgments. This paper is supported by National Natural Science Foundation of China under Grant 61075072, & 91220301, the Program for New Century Excellent Talents in University under Grant NCET-10-0901.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. The MIT Press (1998)
2. Gullapalli, V., Franklin, J., Benbrahim, H.: Acquiring robot skills via reinforcement learning. *IEEE Control Systems* 14(1), 13–24 (1994)
3. Xu, X., Liu, C., Yang, S., Hu, D.: Hierarchical approximate policy iteration with binary-tree state space decomposition. *IEEE Transactions on Neural Networks* 22(12), 1863–1877 (2011)
4. Wang, F.Y., Zhang, H., Liu, D.: Adaptive dynamic programming: An introduction. *IEEE Computational Intelligence Magazine* 4(2), 39–47 (2009)
5. Sutton, R.S.: Generalization in reinforcement learning: successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems* 8, 1038–1044 (1996)
6. Lagoudakis, M.G., Parr, R.: Least-squares policy iteration. *Journal of Machine Learning Research* 4, 1107–1149 (2003)
7. Liu, D., Javaherian, H., Kovalenko, O., Huang, T.: Adaptive critic learning techniques for engine torque and air-fuel ratio control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 38(4), 988–993 (2008)
8. Mahadevan, S.: Representation policy iteration. In: *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (AAAI)*, pp. 372–377 (2005)

A Combined MRSMC/MBC Altitude Controller for a Quad-rotor UAV

Wei Wang¹, Hao Ma¹, and Changyin Sun²

¹ Nanjing University of Information Science & Technology, Nanjing,
Jiangsu P.R. 210044, China
wwcb@nuiist.edu.cn

² Southeast University, Nanjing, Jiangsu P.R. 210096
cysun@seu.edu.cn

Abstract. This paper mainly discusses the altitude control of a quad-rotor UAV. At first, a brief introduction is given to the experimental setup. The altitude information is estimated by the barometer and accelerometer to improve the accuracy. Then a novel altitude controller which combined the Model Reference Sliding Mode Controller (MRSMC) and Memory Based Controller (MBC) is proposed. The experimental results show the good performance of the improved controller.

Keywords: Quad-rotor, Altitude control, MRSMC, MBC.

1 Introduction

In recent years, people have witnessed growing devastating disasters like earthquake, flood and nuclear pollution. In such situations, it will be very useful if a flying robot can access the dangerous or difficult area to monitor the situation with a camera and various sensors. Therefore, the research on UAVs which is believed to play an important role is becoming more and more prevalent.

The quad-rotor, an especially appealing type of UAV, is drawing a rapid increasing attention. Having a wide application in surveillance, search-rescue and information access, the vehicle uses two pairs of counter-rotating, fixed-pitch rotors located at the four corners of the aircraft. The stability and flight control are achieved by changing the rotor speed of propellers [1]. A number of quad-rotor platforms has been developed not only by research groups such as MIT; Heudiasyc; Pennsylvania; Chiba University, but also commercial organizations such as Ascending Technologies; DraganFly Innovations; Microdrones [2-4]. The representative control methods have PID control [5], optimal control [6], sliding model control [7], Precision control [8], nonlinear hierarchical control [9], adaptive regulation control [10] and backstepping control [11].

In this paper, the research object is a quad-rotor UAV that designed by our research group. The attitude control has been completed in previous work and the following sections mainly discuss the stable altitude control. To improve the measurement accuracy and overcome the drifting, the vertical acceleration and barometer sensor

data are supplied to the Kalman filter to estimate the altitude information. Then a novel combined MRSMC/MBC controller is proposed. The MRSMC uses a reference model to give the ideal response of the target value then the sliding mode controller is designed to force the actual output to track the ideal response. When the system error between the output and the target value is within a given small range, the controller is switched to MBC which makes uses of certain history information and the performance is not related to the detail of the system dynamics. Note that, the MBC can track the target value with little error within a few sampling period, so it's usually used together with other control algorithm to avoid too large control input.

2 Description of the Quad-rotor

The overview of our platform is shown in Fig.1. The airframe is made of carbon fiber material, so it provides significant rigidity and light weight. Brushless motors from Scorpion and Pentium-18A Ultra-PWM brushless controllers are selected to power the vehicle. With a three-cell-2100mAh LiPo battery and APC1047 propellers, the specification of the vehicle is shown in Table.1.

The embedded control system applies two STM32F107VCT6 micro controllers to do high frequency process (400Hz attitude control, 50Hz altitude control). The sensors installed are three gyroscopes (ADXRS610), one three-axes accelerometer (ADXL335), one magnetometer (MAG3110) and one barometer (BMP0805). Additionally, a 2.4G wireless SBUS receiver is adopted to receive the control instruction; an XBee-pro is adopted to accomplish the communication between the vehicle and the ground control station. The configuration of the whole system is shown in Fig.2.



Fig. 1. Quad-rotor UAV

Table 1. Specification of quad-rotor UAV

Parameter	Value	Unit
Diameter	500	mm
Height	250	mm
Total mass	1000	g
Maximum lift	800	g
Flight time	15	min

3 Altitude Modeling

The altitude of the vehicle is controlled by the rotational speed of four rotors. As a common sense, the generated thrust F is proportion to the square of the rotation speed n . Considering the hovering movement of a vehicle mass m , the thrust F_0 is

$$F_0 = kn_0^2 = mg / 4 \quad (1)$$

Here, k is a constant coefficient. When the rotation speed is changed to $n = n_0 + \Delta n$, the variation of the thrust ΔF is generated

$$\Delta F = 2kn_0\Delta n + k\Delta n^2 = \sqrt{km}g\Delta n + k\Delta n^2 \quad (2)$$

As a supposition, the Δn is small enough and the Δn^2 can be ignored. The Eq.(2) can be driven to

$$\Delta F = \sqrt{km}g\Delta n = k_1\Delta n \quad (3)$$

In consideration of the delay and the motor load which can be treated as a one order inertia system, the following equation can be obtained:

$$\Delta F(s) = \frac{k_1 k_2}{1 + Ts} u(s) \quad (4)$$

$$a_z = \ddot{L} = (\Delta F C_\theta C_\phi - f_Z) / m \quad (5)$$

L is the altitude, $C_x = \cos x$, ϕ, θ is roll and pitch angle, f_Z is the air-resistance

$$f_Z = k_3 \dot{L} \quad (6)$$

From Eq. (4), Eq. (5) and Eq. (6), we can get the following transfer function

$$G(s) = \frac{Z(s)}{u(s)} = \frac{k_1 k_2 C \theta C \phi}{k_3 s (1 + Ts) (1 + \frac{m}{k_3} s)} \quad (7)$$

Consequently, the complete dynamic state model which governs the altitude MAVs is formed by the following equation.

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -\frac{m + Tk_3}{Tm} & \frac{k_3}{Tm} & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \frac{k_1 k_2 C \theta C \phi}{Tm} \\ 0 \\ 0 \end{bmatrix} u \quad (8)$$

The unknown parameters in the above equation are determined by system identification and the time history response is analyzed in Fig.3 to verify the accuracy of the identified model.

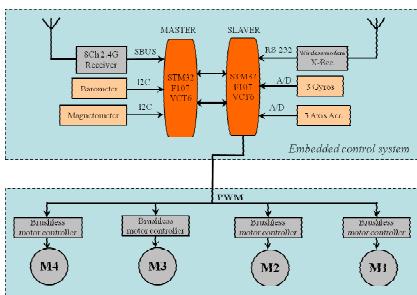


Fig. 2. Configuration of 4 rotors type MAV system

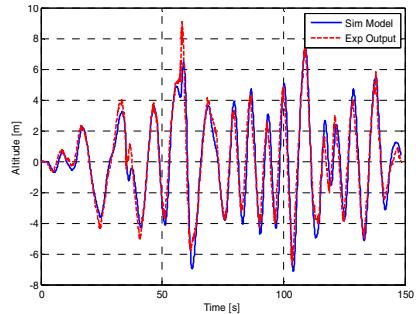


Fig. 3. Cross validation result of altitude model

4 Controller Design

4.1 MRSMC Design

The basic idea of MRSMC is to force the tracking error dynamics between the plant and the reference model states, instead of the plant dynamics, in sliding mode. A brief block diagram of MRSMC is shown in Fig.4. The design of the controller can be divided into the following two steps.

Step1. Design of the reference model

A convenient way to generate a reference model is to specify a model whose states correspond directly to the states of identified model,

$$\begin{cases} \dot{x}_r = A_r x_r + B_r r \\ y_r = C_r x_r \end{cases} \quad (9)$$

with r is the target value. By denoting $C_r = C$, $e = x - x_r$, the error dynamics is obtained by subtracting Eq.(9) from Eq.(8)

$$\dot{e} = A_r e + (A - A_r)x + Bu - B_r r \quad (10)$$

Assuming this system satisfies the following matching condition.

$$A_r - A = BK_1, B_r = BK_2 \quad (11)$$

Then

$$\dot{e} = A_r e - B(K_1 x + K_2 r - u) \quad (12)$$

To ensure the output y_m is settled to r , the DC gain must be adjusted to 1. So

$$K_2 = (-C_r A_r^{-1} B)^{-1} \quad (13)$$

Then B_r can be calculate by

$$B_r = BK_2 = B(-C_r A_r^{-1} B)^{-1} \quad (14)$$

A_r is determined by adjusting in Matlab simulation.

$$A_r = \begin{bmatrix} -a_1 & -a_2 & -a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (15)$$

In this work, the parameters are chosen to be $a_1=6$, $a_2=9.5$, $a_3=5.5$.

Step2. Design of the SMC feedback loop

To strengthen the tracking performance, a new state \mathcal{E}_y that integrates the error e_y is introduced to the system.

$$\dot{\mathcal{E}}_y = y - y_m \quad (16)$$

The extended system can be given as

$$\dot{\mathcal{E}}_s = \begin{bmatrix} \dot{e} \\ \dot{\mathcal{E}}_y \end{bmatrix} = \begin{bmatrix} A_r & 0 \\ C_r & 0 \end{bmatrix} \begin{bmatrix} e \\ \mathcal{E}_y \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u_s = A_s e_s + B_s u_s \quad (17)$$

with $u_s = -(K_1 x + K_2 r - u)$.

For the error system, the sliding surface σ is defined as

$$\sigma = S e_s \quad (18)$$

When the plant is restricted on the sliding surface, $\sigma = \dot{\sigma} = 0$ and equivalent input is derived.

$$u_{eq} = -(S B_s)^{-1} S A_s e_s + K_1 x + K_2 r \quad (19)$$

When u_{eq} is substituted for Eq.(17) as the control input , Eq.(20) is obtained:

$$\dot{e}_s = \left\{ I - B_s (S B_s)^{-1} S \right\} A_s e_s \quad (20)$$

It is known that the system of Eq. (20) becomes stable by stabilizing zeros. To determine the hyper plane S, the optimal control method is adopted and the feedback gain F is selected.

$$F = S = B_s^T P \quad (21)$$

With P is the solution of the Riccati equation.

$$P A_s + A_s^T P - P B_s B_s^T P + Q = 0 \quad (22)$$

In order to maintain the system to be always on the hyper plane, a nonlinear input of sliding mode controller is needed.

$$u_{nl} = K f(\sigma) \quad (23)$$

Where K is switching amplitude and $f(\sigma)$ is a smoothing switching function defined as

$$f(\sigma) = \frac{\sigma}{|\sigma| + \delta} \quad (24)$$

with δ is the weight of smoothing.

Additionally, to generate the stable variables, a Kalman filter is introduced to estimate the unmeasured state conditions. Fig.5 gives the simulation result of MRSMC for step signal.

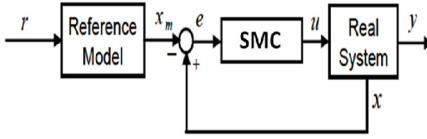


Fig. 4. Block diagram of MRSMC

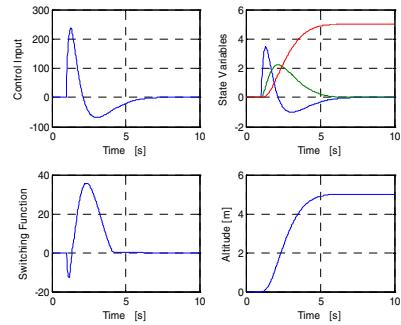


Fig. 5. Simulation result of MRSMC

4.2 MBC Design

The above controller is designed based on the linearized model. It cannot work well in certain instances. Therefore, a memory based controller is introduced to make up the defect.

First, rewrite the differential equation of the altitude dynamic

$$\ddot{L} = -\frac{m + k_3 T}{T m} \dot{L} - \frac{k_3}{T m} \dot{L} + \frac{k_1 k_2 C \theta C \phi}{T m} u = f(\dot{L}, \dot{L}, L) + K u \quad (25)$$

Defining error $e = L - L^*$, the error system can be described as

$$\ddot{e} = f(\dot{L}, \dot{L}, L) + K u - \ddot{L}^* \quad (26)$$

With L^* is the reference model output.

For convenience, a new variable s is introduced

$$s = (q + k_0)^2 e, (k_0 > 0, q = \frac{d}{dt}) \quad (27)$$

The advantage of this treatment is that the original 2-order tracking problem in respect of ϕ is reduced to 1-order stabilization in s .

To substitute the \ddot{e} , we only need to differentiate s once, leading to

$$\dot{s} = (q + k_0)^2 \dot{e} = \ddot{e} + 2k_0 \dot{e} + k_0^2 e \quad (28)$$

Replacing \ddot{e} above by Eq. (26), we get

$$\dot{s} = f(\ddot{L}, \dot{L}, L) - \ddot{L}^* + 2k_0 \dot{e} + k_0^2 e + Ku \quad (29)$$

Now, define the control input as

$$Ku = v + \eta(.) \quad (30)$$

With $\eta(.) = \ddot{L}^* - 2k_0 \dot{e} - k_0^2 e$, and v is a memory-based compensator accounting for system nonlinearities and other disturbances. The closed-loop system becomes

$$\dot{s} = f(\ddot{L}, \dot{L}, L) + v \quad (31)$$

By Euler Differentiation, we get

$$s_{k+1} = s_k + (f_k + v_k) * T \quad (32)$$

Here, T is the sampling period. Now performing backward time-shift in the above equation, we have

$$s_k = s_{k-1} + (f_{k-1} + v_{k-1}) * T \quad (33)$$

Subtracting Eq. (32) from Eq. (33) gives

$$s_{k+1} = 2s_k - s_{k-1} + (f_k + v_k) * T - (f_{k-1} + v_{k-1}) * T \quad (34)$$

Defining v_k as

$$v_k = w_0 v_{k-1} + w_1 s_k + w_2 s_{k-1} \quad (35)$$

with $w_0 = 1$, $w_1 = -\frac{2}{T}$, $w_2 = \frac{1}{T}$. Then Eq. (34) can be rewritten as

$$s_{k+1} = (f_k - f_{k-1})T = T \Delta f_k \quad (36)$$

For a fairly small T , $\Delta f_k \approx T \frac{df}{dt}|t = kT$, so we have

$$|s_{k+1}| \leq T^2 \lambda \quad (37)$$

λ in the above equation is defined as :

$$\lambda = \max\left(\left|\frac{df}{dt}\right| \middle| t = 0, T, 2T, \dots, nT\right) \quad (38)$$

The definition of s can be regarded as a sliding hyper plane with a chattering range of $T^2 \lambda$, when the s is convergent, the tracking error e and its derivative are also convergent.

To summarize, the memory based controller can be generated by subtracting $u_{k-1} = K^{-1}(v_{k-1} + \eta_{k-1})$ from Eq. (30)

$$u_k = u_{k-1} - \frac{2K^{-1}}{T}(\ddot{e}_k + 2k_0\dot{e}_k + k_0^2 e_k) + \frac{K^{-1}}{T}(\ddot{e}_{k-1} + 2k_0\dot{e}_{k-1} + k_0^2 e_{k-1}) + K^{-1}(\eta_k - \eta_{k-1}) \quad (39)$$

with $\eta(.) = \ddot{L}^* - 2k_0\ddot{e} - k_0^2\dot{e}$.

For simplification, $\eta(.)$ can be merged into $f(.)$, so the expression is driven to

$$u_k = u_{k-1} - \frac{2K^{-1}}{T}(\ddot{e}_k + 2k_0\dot{e}_k + k_0^2 e_k) + \frac{K^{-1}}{T}(\ddot{e}_{k-1} + 2k_0\dot{e}_{k-1} + k_0^2 e_{k-1}) \quad (40)$$

5 Experimental Results

In this section, we shall demonstrate that the proposed controller is applicable to the UAV and the outer door experiment is carried out. At first, let the UAV hovering stable for some time with only the attitude controller works. Then the altitude controller is introduced and the current altitude is regarded as the zero. Step and ramp signal are given to the UAV to verify the performance, the experimental result is given in figures below. We can see that the output altitude can track the step target value well in about 5s. The fluctuation is caused by the sensor noise of the barometer. Fig.7 also gives a perfect tracking performance of the ramp signal.

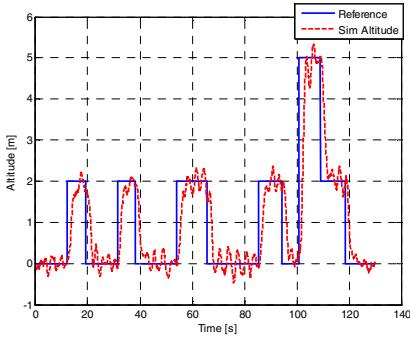


Fig. 6. Experimental result of step signal

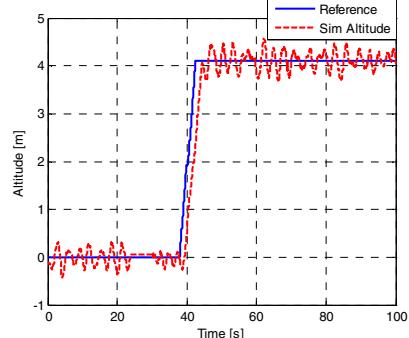


Fig. 7. Experimental result of ramp signal

6 Conclusion

This paper mainly states the altitude control of a quad-rotor UAV, including a brief introduction of hardware setup and the controller design details. To obtain the accurate altitude, the pressure data and vertical acceleration are introduced to the Kalman filter to estimate the altitude information. The controller adopted is a combination of the MRSMC and MBC to enhance the robustness and make up the

model error caused by linearization. The outdoor experiments show that the designed controller can track the target value well and the good stability is also obtained.

References

1. Pebrianti, D., Wang, W., Iwakura, D., et al.: Sliding Mode Controller for Stereo Vision Based Autonomous Flight of Quad-Rotor MAV. *Journal of Robotics and Mechatronics* 23(1), 137 (2011)
2. Achtelik, M., Bachrach, A., He, R., et al.: Stereo vision and laser odometry for autonomous helicopters in GPS-denied indoor environments. In: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, 733219-733219-10 (2009)
3. Bachrach, A.G.: Autonomous flight in unstructured and unknown indoor environments. Massachusetts Institute of Technology (2009)
4. Nonami, K., Kendoul, F., Suzuki, S., et al.: Autonomous Flying Robots: Unmanned Aerial Vehicles and Micro Aerial Vehicles. Springer Publishing Company, Incorporated (2010)
5. Jeong, S.H., Jung, S.: Design and control of a small quad-rotor system under practical limitations. In: 2011 11th International Conference on Control, Automation and Systems (ICCAS), pp. 2011–1163. IEEE (2011)
6. Wang, W., Wang, F., Zhou, Y., et al.: Modeling and embedded autonomous control for quad-rotor MAV. *Applied Mechanics and Materials* 130, 2461–2464 (2012)
7. Wang, W., Nonami, K., Ohira, Y.: Model reference sliding mode control of small helicopter XRB based on vision. *International Journal of Advanced Robotic Systems* 5(3), 235–242 (2008)
8. Hoffmann, G.M., Huang, H., Waslander, S.L., et al.: Precision flight control for a multi-vehicle quadrotor helicopter testbed. *Control Engineering Practice* 19(9), 1023–1036 (2011)
9. Rudin, K., Hua, M.D., Ducard, G., et al.: A robust attitude controller and its application to quadrotor helicopters. In: 18th IFAC World Congress, pp. 10379–10384 (2011)
10. Zeng, W., Xian, B., Diao, C., et al.: Nonlinear adaptive regulation control of a quadrotor unmanned aerial vehicle. In: 2011 IEEE International Conference on Control Applications (CCA), pp. 133–138. IEEE (2011)
11. Lee, D., Ha, C., Zuo, Z.: Backstepping Control of Quadrotor-Type UAVs and Its Application to Teleoperation over the Internet. In: Lee, S., Cho, H., Yoon, K.-J., Lee, J. (eds.) *Intelligent Autonomous Systems 12*. AISC, vol. 194, pp. 217–225. Springer, Heidelberg (2013)

PerGrab: Adapting Grabbing Gesture Recognition for Personalized Non-contact HCI

Tao Li and Ming Li

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{lit,lim}@lamda.nju.edu.cn

Abstract. With recent development of technology, gesture has become a natural way of non-contact human computer interaction. In the literature, to improve user experience of such kind, there exist many works on gesture recognition. However, most works build a universal model for all users, neglecting the fact that different users may have different gesture styles. In this paper, rather than build a universal model for all users, we propose the PerGrab approach by building user-specific model for each user. It is expected that the model can fit users' gesture well, hence leading to better performance. Specifically, given a universal model provided by manufacturers, PerGrab first records user-specific gesture styles by asking users to make some sample gestures, and then employs a personalization step to adapt universal model for the users. Experiments on applications show that PerGrab achieves good performance.

Keywords: Gesture recognition, transfer learning.

1 Introduction

With the development of science and technology, more and more smart equipments appear in our daily life. Traditional touch operation interactions with those devices may be not appropriate in some environments, such as, touching devices is a common way of spread infection in hospitals [1], large screen wall may be too large that some places are out of reach for touching operations. So, non-contact human computer interactions are needed. Naturally, we can use our gestures to communicate with the equipments. In recent years, gesture recognition has found its applications ranging from medical rehabilitation to consumer electronics. In these applications, we can define different gestures for different commands. In the literature, many works [2, 3, 4] have been done.

However, most existing works build a universal model for all users, which neglect the fact that different persons may have different gesture styles. For example, Fig.1 shows the grab and release pictures, where (a) and (b) are from one person and (c) and (d) are from another. It is clearly that, these two persons hold different grab and release gesture styles. If a model build on the data does not cover user specific gesture styles, it will be difficult for such a model to achieve good performance on the corresponding users' gestures. For example, some users

can not communicate with smart TV well because the model offered by the manufacturer is not fit for them. However, it is impossible for the manufacturer to cover all the gesture styles. In consequent, building a model fit for every user is a very challenging task. A better way is that, we just collect a few users' data and use these data to personalize the model offered by the manufacturer. In such case, transfer learning can be used to address this problem. To the best of our knowledge, it is the first work to employ transfer learning in gesture recognition. For simplicity, we consider two gestures recognition problem for example, fist as "left-click" on the mouse and palm as "release left-click".

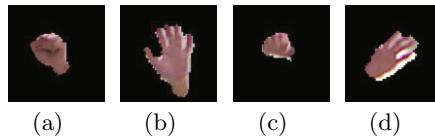


Fig. 1. Different persons have different gesture styles

In this paper, we propose an efficient approach called PerGrab (Personalized Grab), which enables users achieve better performance in hand gesture recognition easily. In the first step, the users are asked to make some sample gestures, such that some user-specific labeled data can be collected. Then based on the collected data, we personalize the universal model offered by the manufacturer by transfer learning. Experiments on the real-world application of PerGrab show that, the performance of PerGrab is very good.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 describes the PerGrab approach including the framework of our system and transfer learning method. In section 4, experiments and results are reported, which are followed by conclusion.

2 Related Work

Vision based hand gesture recognition plays a very important role in human computer interaction (HCI) for it is a non-contact interaction method. In the previous vision based work [2, 5, 6], hands tracking and segmentation are challenging when users are in a complex background. With the development of inexpensive depth sensors, such as Kinect, tracking and segmenting become simpler. Recently, a lot of hand gesture recognition approaches are implemented based on Kincet [4, 7]. They build good performance recognition models for the applications, but they do not consider the fact that different persons may have different gesture styles.

The idea of transfer learning comes from that previous knowledge can help people learn similar new knowledge. Pan and Yang [8] shows transfer learning

finds its utilities in many applications, such as Web-document classification [9], sentiment classification [10], and indoor WiFi location [11, 12]. In these applications, there are only few labeled data for the learning task, while a lot of labeled data available for tasks which are related or similar to the problem. Transfer learning can achieve better performance via exploiting data from different but related tasks.

Due to the success of transfer learning in many applications, we can also treat different persons' hand gesture data as different domains and formulate the hand gesture recognition as a transfer learning problem to get better results.

3 The PerGrab Approach

3.1 The General Framework

The framework of our PerGrab approach is presented in Fig.2. Compared with the traditional universal model methods, PerGrab personalizes the model offered by the manufacturer using a few user's own data, if the user gesture style is different from the manufacturer's.

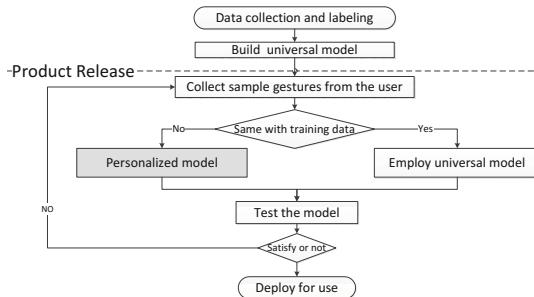


Fig. 2. The PerGrab framework

In the process of PerGrab, before product release, some hand gesture data are collected and labeled, and then a universal model is built. After product release, when users use their equipments for the first time, an initial setting is offered to them. They will be asked to make some sample gestures obeying the setting program to record some users' labeled data. After collecting a few users' data, non-parametric two-sample test Maximum Mean Discrepancy (MMD) [13] will be used to judge whether the customer's gesture style is similar to the training data offered by the manufacturer. If they are similar, the universal model will be directly employed, otherwise transfer learning method will be used to personalize manufacturer's universal model to get a user-specific one. Finally, the users can communicate with the equipment by their hand gestures on user-specific model.

3.2 Adapting Gesture Recognition Model for Personalization

As shown in Fig.2, the "Personalized model" step is the core of PerGrab. Let $x \in R^n$ denote a sample, $y \in \{+1, -1\}$ as the corresponding label. Due to different persons may have different hand gesture styles and the limitation that only a few users' data are available, this personalized model problem can be formulated as a transfer learning problem. Manufacturer offers plenty of labeled source domain data and users offer a few target domain data by initial setting. Therefore, we can personalize the manufacturer's universal model. Set $D_S = \{(x_{s_1}, y_{s_1}), (x_{s_2}, y_{s_2}), \dots, (x_{s_{ns}}, y_{s_{ns}})\}$ as the data for source domain (ns is the source sample size). Moreover, $D_T = \{(x_{t_1}, y_{t_1}), (x_{t_2}, y_{t_2}), \dots, (x_{t_{nt}}, y_{t_{nt}})\}$ donates data for target domain (nt is the target sample size).

Although different persons may have different gesture styles, their gestures should be similar, and parameters of each user's optimal recognition model are highly related. Specifically, we assume that parameters of each model can be spilt into two parts, one common part w_0 and the specific part w_t (t is the related model number). The problem is formulated as transfer learning, where manufacturer's plenty of labeled data are viewed as source domain and users' a few labeled data are recorded as target domain. Set w_s and w_t as source domain's parameters and target domain's parameters, which can be presented as follows:

$$w_s = w_0 + v_s \quad \text{and} \quad w_t = w_0 + v_t \quad (1)$$

where w_0 is the common part of the two domains, v_s is the source-specific part and v_t is the target-specific part.

We formulate the problem as follows:

$$\begin{aligned} & \min_{w_0, v_t, \xi_{t_i}} \frac{1}{2} \|w_0\|^2 + \frac{C_1}{2} \|v_s\|^2 + \frac{C_2}{2} \|v_t\|^2 + C_3 \sum_{i=1}^{n_s} \xi_{s_i} + C_4 \sum_{i=1}^{n_t} \xi_{t_i} \\ & \text{s.t. } y_{s_i}(w_0 + v_s) \cdot x_{s_i} \geq 1 - \xi_{s_i}, \xi_{s_i} \geq 0 \\ & \quad y_{t_i}(w_0 + v_t) \cdot x_{t_i} \geq 1 - \xi_{t_i}, \xi_{t_i} \geq 0 \end{aligned} \quad (2)$$

The first three parts of equation (2) are regularization terms. The first term is common part parameters, the second is source-specific part and the third is target-specific part. We use C_1 and C_2 to adjust their relationship. The last two parts are the sum of hinge loss of source domain and target domain. Different from the multi-task learning [14] that care about all the tasks' performance, we just focus on the target domain's performance in the transfer learning. So we give a higher value to C_4 than C_3 .

We rewrite the formulation above into the dual form:

$$\min_{\substack{0 \leq \lambda_S \leq C_3 \\ 0 \leq \lambda_T \leq C_4}} \frac{1}{2} \begin{pmatrix} \lambda_S \\ \lambda_T \end{pmatrix}^T \begin{pmatrix} \left(1 + \frac{1}{C_1}\right) A_S & A_{ST} \\ A_{TS} & \left(1 + \frac{1}{C_2}\right) A_T \end{pmatrix} \begin{pmatrix} \lambda_S \\ \lambda_T \end{pmatrix} + \mathbf{1}^T \begin{pmatrix} \lambda_S \\ \lambda_T \end{pmatrix} \quad (3)$$

$$\text{s.t. } \begin{pmatrix} y_S^T & \mathbf{0} \\ \mathbf{0} & y_T^T \end{pmatrix} \begin{pmatrix} \lambda_S \\ \lambda_T \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

where C_1, C_2, C_3, C_4 are the same as equation (2). A_S is the kernel for the source domain, with element $A_{S_{i,j}} = K(x_{s_i}, x_{s_j})$, $x_{s_i}, x_{s_j} \in D_S$. A_{ST} is the kernel for the source and target domain, with element $A_{ST_{i,j}} = K(x_{s_i}, x_{t_j})$, $x_{s_i} \in D_S, x_{t_j} \in D_T$. A_T and A_{TS} are similar. The kernel can be linear, polynomial or others. In this paper, we use the linear kernel. Mosek is employed to deal with this quadratic programming problem in seconds.

4 Experiments

There are 8 persons taking part in the experiment, and each person offers some data using the sampling program we design. The program tells each participant to do some grab and release gestures, and records his/her hand areas in 100×100 pixels. We record 2000 data for each person, 1000 positive (fists) and 1000 negative (palms).

Features used in this paper are Histogram oriented Gradient (HoG) in 4000 dimensions and Gist in 512 dimensions. The baseline methods are SVM and Random Forest, which are widely used in computer vision applications.

Three experiments are conducted in this section. Firstly, we demonstrate different persons may have different gesture styles. Secondly, model personalization by adapting the universal can get good performance. Finally, we show PerGarb gets better performance when the number of training persons increases. Average F1 measure over 50 rounds is used for fist.

4.1 Experiment 1: Universal Model Does Not Work for Every User

We want to prove different persons may have different gesture styles. We think if the universal model does not cover someone's style, it will not work for him/her. we conduct experiments as follows. One person is assigned as training data, randomly sampling 1000 samples (500 fists + 500 palms) to train a model, then test the model on all the 8 persons' data and get 8 results. The testing data consist of 1000 samples (500 fists + 500 palms), the training data and testing data are without overlapping when they come from the same person. We can get 8 groups of results by changing the training person.

Table 1. Average Fist F1 measure of each pair of classifiers and features

	SVM(Polynomial)	SVM(Linear)	SVM(RBF)	Random Forest
HoG	.937 ± .047	.932 ± .048	.932 ± .048	.907 ± .062
Gist	.916 ± .065	.915 ± .061	.921 ± .057	.912 ± .050

We conduct the experiments of all pairs of features and baseline methods with cross-validation to get the each best results. The average results are shown in Table 1. We find the pair of SVM with polynomial kernel and HoG feature gets the best performance, so we choose this pair on behalf to conduct our following

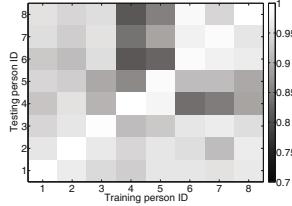


Fig.3. Results of the model of SVM(Polynomial) + HoG

experiments. In details, the 8×8 result matrix of this pair of classifier and feature is shown in Fig.3, while the other pairs get the similar results.

Fig.3 shows the results of the model on HoG feature and SVM with polynomial kernel. We can see that each model trained by one person's data can get good performance on some persons, but get bad performance on some other persons' data. For example, when the training person ID = 3, the model performs good on Testing person ID = 6,7,8, but bad on Testing person ID = 4,5. The similar results appear when other pairs of features and classifiers are employed. So, It is clear that different persons may have different gesture styles, and a universal model can't satisfy all the users.

4.2 Experiment 2: Adapting Universal Model Helps

Each training person in the Experiment 1 is selected as source domain, with 950 samples (475 fists + 475 palms) sampled, and its testing persons are selected as target domains, offering 50 target domain samples (25 fists + 25 palms) and 1000 testing samples (500 fists + 500 palms). Parameters of PerGrab are set as $C_1 = C_2 = C_3 = 1$ and $C_4 = 10$. At the same time, we also simply combine the source and target data as the new training data under the unified distribution assumption and build the model using SVM. The two 8×8 result matrixes are shown in Fig.4.

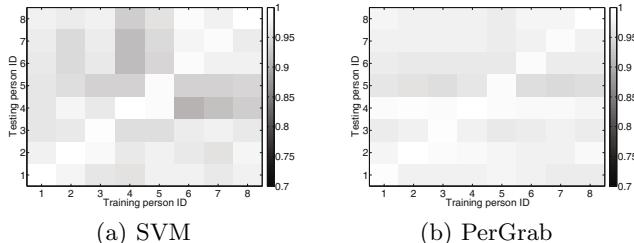


Fig.4. The results of SVM and PerGrab trained by source and target domain data

Fig.4 shows results of PerGrab and its compared method. After sign test by t-test with significance level 0.05, PerGrab gets 40 wins, 19 ties and 5 loses, which is significantly better than the results shown in Fig.3. From the results, we can say transfer learning really helps in the hand gesture recognition application.

4.3 Experiment 3: When More Persons Are Added to Training Set

In this subsection, we conduct experiment in conditions where more source persons' data are available. The experiment is conducted with source domain person number ranging from 1 to 5. For the fair of the experiment, the sum of all the source data size and target data size is set at 1000. Assume we have N source persons' data, we can get N transfer models for each pair of source person and target person. When target person's testing data come, each testing datum will get N predict labels by N transfer models. Finally, all the predicted labels are ensembled by majority voting to get the final label prediction. The results are shown in Fig.5.

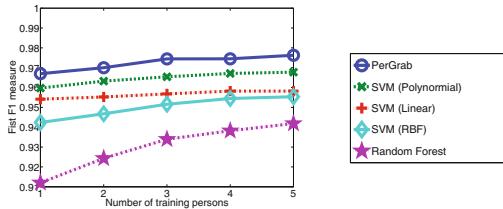


Fig. 5. The results of PerGrab and the comparison methods as the number of training persons increases

Fig.5 shows that as the number of training person increases, the performance of PerGrab and comparison methods become better. Moreover, our PerGrab gets the best performance among all the methods.

5 Conclusion

In this paper, we discuss the problem that different persons may have different gesture styles, due to which, universal model is not able to give satisfactory performance for all users. And we introduce the PerGrab approach, which adapts the universal model offered by manufacturer to the user-specific model. PerGrab just needs user to do a little work to collect a few data as user-specific data, which are used to personalize the model offered by the manufacturer. The experiments show PerGrab performs better than universal distribution formulation.

In the current paper, we only consider the grab and release gestures. The framework of PerGrab can also be applied to other gestures recognition applications, such as gestures in sign language. In the future, we will continue our research on other complex gestures recognition issues.

Acknowledgments. The work is supported by the National Science Foundation of China (No. 61272217), and National Social Science Funds of China (No. 11AZD121) and the 2013 State Grid Research Project.

References

- [1] Schultz, M., Gill, J., Zubairi, S., Huber, R., Gordin, F.: Bacterial contamination of computer keyboards in a teaching hospital. *Infection Control and Hospital Epidemiology* 24(4), 302–303 (2003)
- [2] Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture recognition using multi-scale color feature, hierarchical models and particle filtering. In: Proceedings of the 5th Face and Gesture, pp. 423–428 (2002)
- [3] Zhang, X., Chen, X., Wang, W., Yang, J., Vuokko, L., Wang, K.: Hand gesture recognition and virtual game control based on 3d accelerometer and emg sensors. In: Proceedings of the 14th IUI, pp. 401–406 (2009)
- [4] Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover’s distance with a commodity depth camera. In: Proceedings of the 19th ACM MM, pp. 1093–1096 (2011)
- [5] Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.C.: Filtering using a tree-based estimator. In: Proceedings of the 9th ICCV, pp. 1063–1070 (2003)
- [6] Wang, C.C., Wang, K.C.: Hand posture recognition using adaboost with sift for human robot interaction. In: Lee, S., Suh, I.H., Kim, M.S. (eds.) *Recent Progress in Robotics: Viable Robotic Service to Human.* LNCIS, vol. 370, pp. 317–329. Springer, Heidelberg (2008)
- [7] Li, H., Yang, L., Wu, X., Xu, S., Wang, Y.: Static hand gesture recognition based on hog with kinect. In: Proceedings of the 4th IHMSC, pp. 271–273 (2012)
- [8] Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10), 1345–1359 (2010)
- [9] Dai, W., Yang, Q., Xue, G., Yu, Y.: Boosting for transfer learning. In: Proceedings of the 24th ICML, pp. 193–200 (2007)
- [10] Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th ACL, pp. 432–439 (2007)
- [11] Zheng, V., Pan, S.J., Yang, Q., Pan, J.: Transferring multi-device localization models using latent multi-task learning. In: Proceedings of the 23rd AAAI, pp. 1427–1432 (2008)
- [12] Pan, S., Zheng, V.W., Yang, Q., Hu, D.H.: Transfer learning for wifi-based indoor localization. In: Proceedings of the 23th AAAI (2008)
- [13] Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. In: Proceedings of the 19th NIPS, pp. 513–520 (2007)
- [14] Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: Proceedings of the 10th KDD, pp. 109–117 (2004)

An Automatic MSRM Method with a Feedback Based on Shape Information for Auroral Oval Segmentation

Hui Liu, Xinbo Gao, Bing Han, and Xi Yang

School of Electronic Engineering, Xidian University, Xi'an 710071, China
liuhuileucky@yeah.net, xbgao@mail.xidian.edu.cn,
bhan@xidian.edu.cn, xi.yang.xidian@gmail.com

Abstract. Auroral oval segmentation is of great significance to the study of auroral activities. In this paper, we propose an automatic maximal similarity based region merging (MSRM) method with a feedback based on shape information. Firstly, K -means method is employed to mark auroral oval points and background points, thus guiding the process of MRSR to obtain the initial segmentation result. Then the direct least-square ellipse fitting method is used to fit an ellipse on the initial boundary and points in the fitted ellipse are set as adjusted markers of auroral oval. Finally, the MRSR mechanism is used again to get the final segmentation result. Experimental results show that the proposed algorithm obtains a good performance.

Keywords: Auroral oval segmentation, maximal similarity, region merging, direct least-square ellipse fitting.

1 Introduction

Aurora is a natural light phenomenon that usually appears in the sky of high latitude regions, which is caused by the collision of solar wind and magnetospheric particles [1]. The ultraviolet imager (UVI) image of aurora is one of the means to reflect auroral activity. From the spatial morphology of auroral oval in UVI image, we can infer magnetospheric activity indices effectively. Thus, research on a relatively accurate segmentation of auroral oval can contribute to the study of magnetospheric activity.

Five typical UVI images of auroral oval are shown in Fig. 1. We can see that there are a lot of noise and interference in the image and most images are of low contrast. So it's difficult to segment the auroral oval in UVI images.

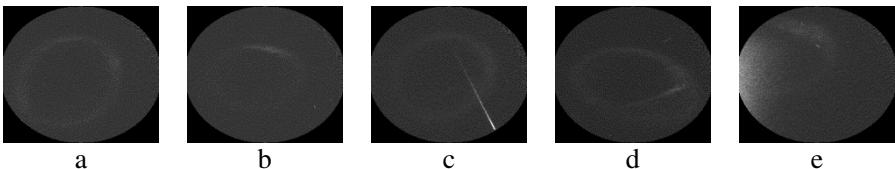


Fig. 1. Typical UVI images of auroral oval: (a) noisy image; (b) low contrast image; (c) image with meteor; (d) θ -aurora image; (e) dayglow contaminated image.

Most of the existing methods for auroral oval segmentation are pixel-based, such as adaptive minimum error thresholding (AMET) [2] and histogram-based K -means (HKM) [3] methods. However, these pixel-based methods cannot get a complete auroral oval boundary in practice. To overcome this problem, a shape-based auroral oval segmentation method driven by linear least-squares randomized Hough transform (LLS-RHT) [4] is proposed subsequently. In LLS-RHT, two ellipses are obtained as the inner and outer auroral oval boundaries, respectively. The boundaries gotten by LLS-RHT are complete, but the smooth inner boundary conflicts with the fact that there are extended bulges on the actual inner boundary of auroral oval. As the area within the inner boundary of auroral oval is associated with magnetic flux, an accurate inner boundary is more helpful to do further research on magnetic flux. Therefore, the LLS-RHT method is not suitable in this case and a new precise segmentation method is necessary.

To get a more accurate and complete boundary of auroral oval, an automatic maximal similarity based region merging (MSRM) method with a feedback based on shape information for auroral oval segmentation is presented in this paper. The MSLR [5] method is adaptive to the image content and it does not need to set the similarity threshold in advance. It can extract the object contour under the guidance of object and background markers. If there are object markers in the low contrast region of auroral oval, a complete boundary of auroral oval can be gotten by this method. But the original MSLR algorithm is not robust because of the using of manual markers. What's more, it's tedious to mark the auroral oval and the background manually over a series of images. To address the aforementioned problems, novel automatic markers are used in this paper. The new auroral oval segmentation method is a region-based method with shape information. The MSLR mechanism is used to locate the auroral oval with a fitted ellipse obtained automatically as object markers, by which a complete auroral oval boundary can be obtained from the low contrast UVI image robustly. In addition, the inner boundary of the auroral oval is relatively accurate.

This paper is organized as follows. Section 2 introduces the MSLR method. The proposed auroral oval segmentation method is given in Section 3. Section 4 presents experimental results and analysis. The conclusion of the whole paper is in Section 5.

2 Maximal Similarity Based Region Merging

In this section, the maximal similarity based region merging (MSRM) algorithm is applied to separate the object from the background with the help of markers. When this method is used, the markers must be given by strokes drawn firstly and the markers must be located in the object and the background separately. As the segmentation result is affected by the location of markers and different people will give different markers, the algorithm with manual markers is not robust. In this paper, we adopt an automatic method to make markers. The watershed method is used to obtain the initial segmentation as the input of MSLR algorithm.

2.1 Similarity Measure

Information of edge, texture, color, shape and size can be used to describe a region. Color histogram is better than other descriptors in region merging based segmentation, because the color histogram of different regions from the same object will be highly similar [5]. Similarly, for UVI aurora images, we use gray level histogram to describe a region because UVI images are gray images without any color information.

Let \mathbf{H}_R be the histogram of a region R . We use the Bhattacharyya coefficient $\rho(R, Q)$ as the similarity measure between region R and region Q .

$$\rho(R, Q) = \sum_{i=1}^{256} \sqrt{\mathbf{H}_R^i \bullet \mathbf{H}_Q^i} , \quad (1)$$

where \mathbf{H}_R^i and \mathbf{H}_Q^i are the i th element of gray level histograms in region R and region Q , respectively. There are similar gradation histograms in regions with the same object, so the similarity measure mentioned above will be high between them.

2.2 Maximal Similarity Based Merging Rule

Suppose region Q is adjacent to region R . Q has a set of adjacent regions which can be described as $\bar{S}_Q = \{S_i^Q\}_{i=1,2,\dots,q}$. Of course, $R \in \bar{S}_Q$. If $\rho(Q, R) = \max_{i=1,2,\dots,q} \rho(Q, S_i^Q)$, region R and region Q are merged. The rule means that if the similarity between R and Q is the maximum value among the similarities between Q and \bar{S}_Q , region R and Q will be merged.

3 The Proposed Auroral Oval Segmentation Method

Different from the existing segmentation methods for auroral oval, our method is a region-based method with shape information. The MSRM mechanism is used to locate the auroral oval with an ellipse obtained by direct least-square fitting method [6] as object markers. We can get a more accurate and complete boundary of the auroral oval by the proposed method robustly. The proposed method includes five main steps, which are shown in Fig.2.

A detailed description of each step is given as follows.

Step 1. Get the initial segmentation regions. The original image is $\mathbf{I}_{\text{input}}$. To get the edge information of the image, we use the gradient image obtained by Sobel operator as an input image \mathbf{G} of watershed algorithm. \mathbf{G} is

$$\mathbf{G} = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2} , \quad (2)$$

where \mathbf{G}_x and \mathbf{G}_y are the gradients of the horizontal and vertical directions calculated by Sobel operator, respectively. The initial segmentation regions \mathbf{F} are gotten by using watershed algorithm.

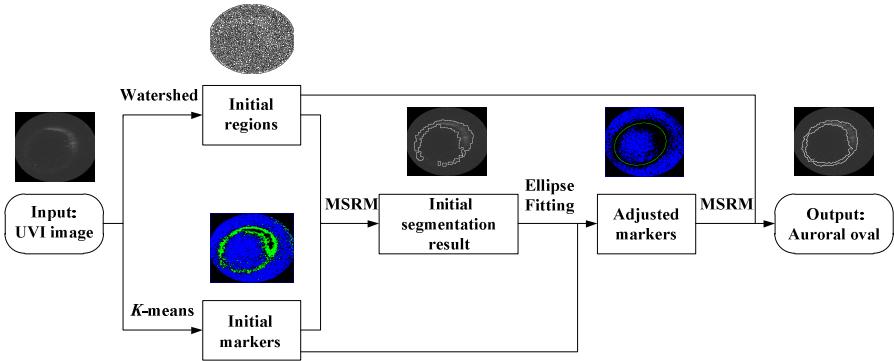


Fig. 2. Diagram of the proposed auroral oval segmentation method

Step 2. Get the initial markers. Based on the characteristics of UVI image, we use the K -means algorithm to cluster the gray level of the image into three parts $\mathbf{P} = \{P_1, P_2, P_3\}$ by minimizing the within-cluster sum of squares

$$\arg \min_{\mathbf{P}} \sum_{i=1}^3 \sum_{x_j \in P_i} \|x_j - c_i\|^2 , \quad (3)$$

where c_i is the center of the points in P_i . The three clustering centers from small to large roughly respect the average gray values of the invalid imaging area, the background and the auroral oval in UVI images separately. We set the points where those gray values are around the middle clustering center and the points in the edge of the whole image as the background markers. The points with gray values around the largest clustering center are regarded as the object markers, that is to say, the auroral oval markers.

Step 3. Use the MSRM algorithm to get the initial segmentation result. We employ \mathbf{M}_b , \mathbf{M}_o and \mathbf{M}_n to denote the sets of regions with background markers, regions with object markers and regions with no markers in \mathbf{F} separately. The merging process is divided into two stages. In the first stage, we use the maximal similarity based merging rule to merge \mathbf{M}_o with its adjacent regions. For each $O \in \mathbf{M}_o$, the set of its adjacent regions is $\bar{S}_O = \{A_i\}_{i=1,2,\dots,r}$. Then for each A_i , $A_i \notin \mathbf{M}_o$ and $A_i \notin \mathbf{M}_b$, we form the set of its adjacent regions $\bar{S}_{A_i} = \{S_j^{A_i}\}_{j=1,2,\dots,k}$. If O and A_i satisfy the rule

$$\rho(A_i, O) = \max_{j=1,2,\dots,k} \rho(A_i, S_j^{A_i}) , \quad (4)$$

O and A_i will be merged. The procedure is executed iteratively until no merging regions can be found. The sets \mathbf{M}_n and \mathbf{M}_o will be updated during each iteration. In the second stage, the regions with no markers merge with their adjacent no-marker regions under the maximal similarity based merging rule. For each $N \in \mathbf{M}_n$, we find the set of its adjacent regions $\bar{S}_N = \{H_i\}_{i=1,2,\dots,p}$. Then for each H_i , $H_i \notin \mathbf{M}_o$ and

$H_i \notin \mathbf{M}_b$, the set of its adjacent regions is $\bar{S}_{H_i} = \{S_j^{H_i}\}_{j=1,2,\dots,q}$. If region N and region H_i satisfy the rule

$$\rho(H_i, N) = \max_{j=1,2,\dots,q} \rho(H_i, S_j^{H_i}) , \quad (5)$$

then N and H_i will be merged. This stage is also implemented iteratively until no merging regions can be found. The above two stages will be implemented repeatedly. The two-stage iterative process stops when new merging regions will not be found. Then we get the initial segmentation image I_{initial} .

Step 4. Get the adjusted markers by ellipse fitting. For the low contrast UVI image, it is hard to get an absolute complete boundary by the steps above. While, from the view of space, the auroral oval has an annular ring shape [7]. Considering a priori knowledge of shape, we can improve our method to get a complete boundary of auroral oval. An ellipse is fitted on the boundary of the initial segmentation auroral oval by using the direct least-square ellipse fitting algorithm [6]. The points in the ellipse are set as adjusted object markers. Then a complete boundary of auroral oval can be obtained by MSRM.

Step 5. Use the MSRM algorithm again to get the final segmentation image I_{output} . Since we set more points in the low contrast region of UVI image as object markers by ellipse fitting method, it is possible to get a complete auroral oval boundary.

4 Experimental Results and Analysis

The experimental results of the proposed method are presented in this section. 30 UVI images of 228×200 provided by the Polar Ultraviolet Imager instrument are used to evaluate our algorithm. All experiments are performed using Matlab R2010b on a Windows 7 platform with an Intel Core i3 CPU at 2.93GHz and 2GB memory.

In order to demonstrate the effectiveness and robustness of our proposed method, we compare the experimental results of our method with AMET [2], HKM [3] and LLS-RHT [4] methods. We label the auroral oval in 30 UVI images firstly. The images labeled manually are ground truth. Then the performance of each method is considered versus the ground truth.

4.1 Subjective Evaluation

Fig.3 shows the segmentation results obtained by the proposed method and the three methods stated above. It can be seen that the pixel-based methods AMET and HKM cannot get a complete boundary of auroral oval in some low contrast images. The LLS-RHT method can give a complete boundary, but the boundary is inaccurate, especially the inner boundary. The performance of our method is good because a relatively complete and accurate boundary is obtained.

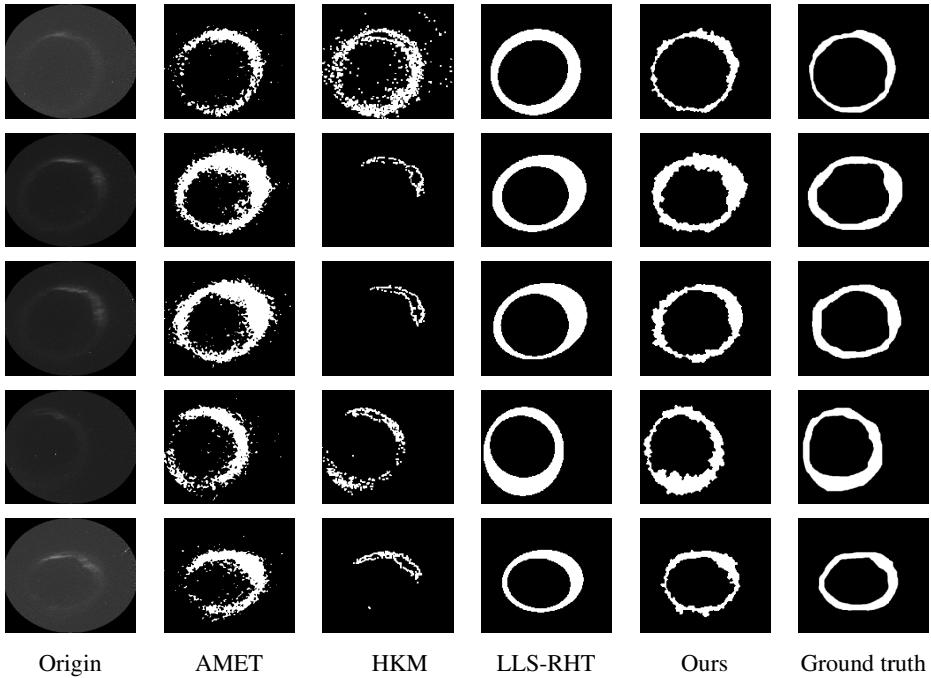


Fig. 3. The segmentation results of different methods. The first column is original image, the second column is the result obtained by AMET method, the third column is the result obtained by HKM method, the fourth column is the result obtained by LLS-RHT method, the fifth column is the result obtained by the proposed method and the last column is ground truth result.

4.2 Objective Evaluation

In this paper, we use two metrics based on the number of the points in the auroral oval to evaluate segmentation accuracy. T and A are the regions obtained by ground truth and other methods. The recall P_r and the false alarm P_e [8] are given by

$$P_r = \frac{N_{T \cap A}}{N_T} , \quad (6)$$

$$P_e = \frac{N_{C_A(T \cap A)}}{N_A} , \quad (7)$$

where $N_{T \cap A}$ is the number of the points both in region T and region A . Similarly, N_T means the number of points in region T and N_A means the number of points in region A . The number of points in region A but not in region T is denoted as $N_{C_A(T \cap A)}$. P_r and P_e of our method and other three methods mentioned above are shown in Fig.4 and Fig.5. From Fig.4 we can see that the proposed method has a high

P_r following LLS-RHT method. As the segmentation region obtained by LLS-RHT method is large, there is much overlap with the region by ground truth. So the P_r obtained by LLS-RHT is higher than other methods. But in Fig.5, the P_e of LLS-RHT is higher than the proposed method. The average P_e of the proposed method is lower than other three algorithms. The P_e of HKM method for some images is low because the segmentation region gotten by HKM method is so small that it only includes the points with a high intensity in the auroral oval and a small number of points outside the auroral oval. These experimental results demonstrate that the proposed method has a good performance on auroral oval segmentation with a high recall and a low false alarm.

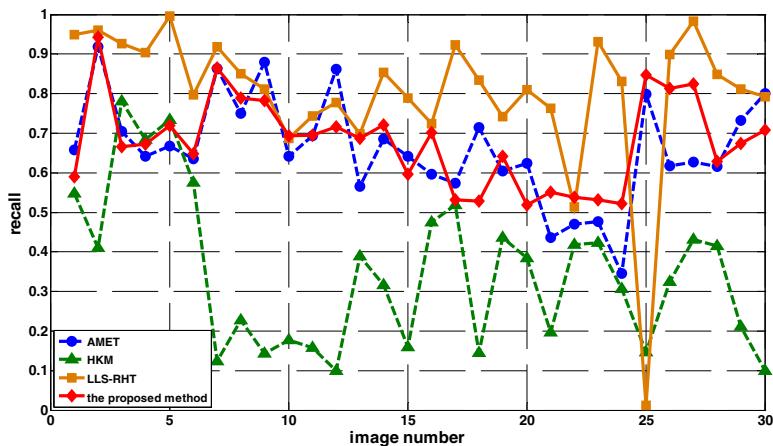


Fig. 4. The recall of the four methods

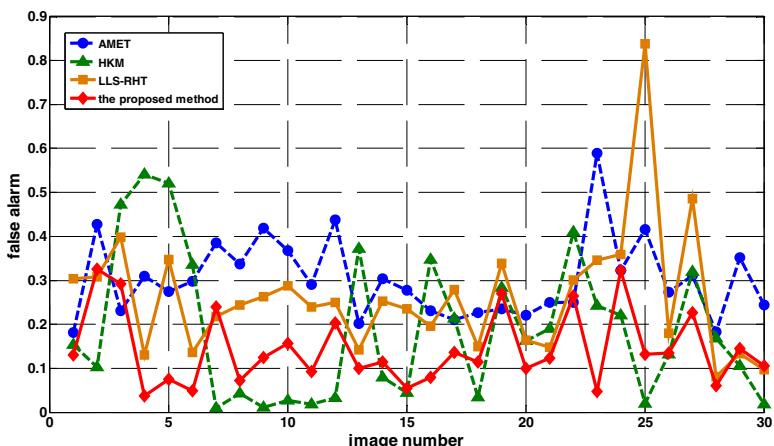


Fig. 5. The false alarm of the four methods

5 Conclusion

In this paper, an automatic MSRM method with a feedback based on shape information for auroral oval segmentation is presented. Experimental results demonstrate that our method can obtain a good performance compared with the existing methods. The MSRM mechanism is used to locate the auroral oval with a fitted ellipse obtained automatically as object markers, by which a complete and precise auroral oval boundary can be obtained from the low contrast UVI image robustly.

Since the ellipse fitting algorithm is used in the proposed method, the method cannot be fully applied to the segmentation of gap oval images, and this is exactly our future work.

Acknowledgements. This research is supported by the National Natural Science Foundation of China (41031064; 60902082), the Shaanxi Province Natural Science Fundamental Research Funded Projects (2011JQ8019), the Special Scientific Research of Marine Public Welfare Industry (201005017), the Basic Foundation for Scientific Research, the Fundamental Research Funds for the Central Universities (K5051302008, K5051202048) and the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

- Germany, G.A., Parks, G.K., Ranganath, H., Elsen, R., Richards, P.G., Swift, W., Spann, J.F., Brittnacher, M.: Analysis of auroral morphology: Substorm precursor and onset on January 10, 1997. *Geophysical Research Letters* 25(15), 3043–3046 (1998)
- Li, X., Ramachandran, R., He, M., Movva, S., Rushing, J., Graves, S., Lyatsky, W., Tan, A., Germany, G.: Comparing different thresholding algorithms for segmenting auroras. In: *Proceedings of the International Conference on Information Technology: Coding and Computing*, Las Vegas, vol. 2, pp. 594–601 (2004)
- Hung, C.C., Germany, G.: K-means and iterative selection algorithms in image segmentation. *IEEE Southeastcon (Session 1: Software Development)* (2003)
- Cao, C.-G., Newman, T.S., Germany, G.A.: New shape-based auroral oval segmentation driven by LLS-RHT. *Pattern Recognition* 42(5), 607–618 (2009)
- Ning, J.-F., Zhang, L., Zhang, D., Wu, C.-K.: Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition* 43(2), 445–456 (2010)
- Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(5), 476–480 (1999)
- Hones, E.W., Craven, J.D., Frank, L.A., Evans, D.S., Newell, P.T.: The horse-collar aurora: A frequent pattern of the aurora in quiet times. *Geophysical Research Letter* 16(1), 37–40 (1989)
- Wang, Y., Hu, B.-G.: Derivations of Normalized Mutual Information in Binary Classifications. In: *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, pp. 155–163 (2009)

An Improved ELM Algorithm Based on EM-ELM and Ridge Regression

Haigang Zhang, Sen Zhang^{*}, and Yixin Yin

School of Automation, University of Science and Technology Beijing,
Beijing, China 100083
zhangsen@ustb.edu.cn

Abstract. Although the extreme learning machine(ELM)algorithm is applied in training single hidden layer feedforward neural networks without manual intervention and its fast training speed has been recognized nowadays, there are still several problems need to be solved in this algorithm. In this paper, we propose a new improved ELM algorithm—Error Minimized and Ridge Regression ELM(ER-ELM) derived from how to select the number of hidden nodes and the way to handle with the multicollinearity. We compare the ER-ELM algorithm with other relative algorithms such as ELM et al. Simulation results show that the ER-ELM algorithm has better stability and generalization performance than the other algorithms.

Keywords: Extreme Learning Machine,Ridge Regression, Single hidden layer neural network.

1 Introduction

In the last few years, feedforward neural networks have received very wide range of applications[1,2]. However, its applications encountered a lot of restrictions such as slow training speed and local minima results. For these two shortcomings, Huang proposed extreme learning machine(ELM) algorithm[3]. ELM is a novel single hidden layer feedforward neural network where the input weights and the bias of hidden nodes are generated randomly without tuning and the output weights is determined analytically.

Although ELM algorithm shows the fast training speed and good generalization property, it still can be further improved in the following aspects: (1) how to select the numbers of hidden nodes, (2) how to deal with the multicollinearity problems. For the first problem, [5]proposed a new method named EM-ELM to select the number of hidden nodes based on the minimal error. P-ELM algorithm proposed in [6] and OP-ELM algorithm proposed in[7] put forward the method to select proper hidden layers at different angles. Some improvements based on PCA [8] and Ridge Regression[9] have been applied to many practical multicollinear problems. This paper presents a new ELM algorithm combined

* Corresponding author.

EM-ELM with Ridge Regression (RR) named ER-ELM, where EM-ELM is applied to select the hidden nodes' number while RR plays a significant role in handling with the multicollinearity.

This paper is organized as follows: the improved ELM algorithm is proposed in section 2. Simulative experiments are given in section 3. Section 4 summarizes the conclusions of this paper.

2 Improved ER-ELM Algorithm

2.1 The Theory of Extreme Learning Machine

Suppose there are N arbitrary samples (x_i, t_i) , where $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n$ denotes an n -dimensional feature of the i th sample and $t_i = [t_{i1}, t_{i2}, \dots, t_{im}] \in R^m$ denotes the target vector. The mathematical model of SLFNs with \tilde{N} hidden nodes is as follows:

$$\sum_{i=1}^{\tilde{N}} \beta_i g_i(x_k) = \sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_k + b_i) = t_k, \quad k = 1, 2, \dots, N \quad (1)$$

where w is $\tilde{N} \times n$ input weight matrix connecting the hidden and the input nodes, β is $\tilde{N} \times m$ output weight matrix connecting the hidden and the output nodes, b is $\tilde{N} \times 1$ bias of hidden layer nodes. And $w_i \cdot x_k$ denotes the inner product of w_i and x_k . t is $N \times 1$ output vector under this model with activation function $g(x)$.

The above n equations can be written in matrix form as

$$H\beta = T \quad (2)$$

$$\text{where } H(W, B, X) = \begin{pmatrix} g(w_1 \cdot x_1 + b_1) & \dots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \dots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{pmatrix}_{N \times \tilde{N}}$$

and β, T

is defined before.

According to the theory of Least Square, the output weight β can be estimated as

$$\hat{\beta} = H^+ T \quad (3)$$

where H^+ is the Moore-Penrose generalized inverse of H .

There are several methods to calculate the Moore-Penrose generalized inverse. Here singular value decomposition(SVD) method is widely used, where $H^+ = (H^T H)^{-1} H^T$. So

$$\hat{\beta} = (H^T H)^{-1} H^T T \quad (4)$$

The above is the theory of ordinary ELM algorithm. Some properties of the solution are discussed as follows [10].

Considering the real complex environment, the model (3) should be modified:

$$T = H\beta + \varepsilon \quad (5)$$

where $\varepsilon \in N(0, \sigma^2)$ denotes white gaussian noise. The solution is:

$$\hat{\beta} = (H^T H)^{-1} H^T T = \frac{\sum_{i=1}^{\tilde{N}} H_i^T (H_i \beta_i + \varepsilon_i)}{\sum_{i=1}^{\tilde{N}} H_i^2} = \beta + \frac{\sum_{i=1}^{\tilde{N}} H_i^T \varepsilon_i}{\sum_{i=1}^{\tilde{N}} H_i^2} \quad (6)$$

$$(1) \quad E(\hat{\beta}) = \beta \quad (7)$$

$$(2) \quad V(\hat{\beta}) = E(\hat{\beta}^2) - E(\hat{\beta})^2 = \frac{\sigma^2}{\sum_{i=1}^{\tilde{N}} H_i^2} = \sigma^2 \sum_{i=1}^{\tilde{N}} \frac{1}{\lambda_i} \quad (8)$$

where λ_i is the i th eigenvalue of $H^T H$.

$$(3) \quad MSE(\hat{\beta}) = \frac{1}{\tilde{N}} E[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] = \frac{1}{\tilde{N}} \frac{\sigma^2}{\sum_{i=1}^{\tilde{N}} H_i^2} = \frac{1}{\tilde{N}} \sigma^2 \sum_{i=1}^{\tilde{N}} \frac{1}{\lambda_i} \quad (9)$$

Suppose $H^T H$ is not always nonsingular, it means: when the H matrix is multicollinear, some eigenvalues will tend to zero, while $V(\hat{\beta})$ and $MSE(\hat{\beta})$ will become larger which will affect its stability and generalization.

2.2 To Overcome Multicollinearity Based on RR

Ridge Regression, which was proposed by Horel and Kennard in 1970, pluses a small positive number on the main diagonal of the design matrix. Considering ELM algorithm, the estimation of $\hat{\beta}^*$ can be obtained by employing the following formula:

$$\hat{\beta}^* = (H^T H + kI)^{-1} H^T T \quad (10)$$

where k is the ridge parameter. Superficially, it is possible to get the inversion of design matrix $H^T H$. The following is the new results of (7), (8), (9).

$$\hat{\beta}^* = \frac{\sum_{i=1}^{\tilde{N}} H_i^T T_i}{\sum_{i=1}^{\tilde{N}} H_i^2 + k} = \frac{\sum_{i=1}^{\tilde{N}} H_i^T (H_i \beta_i + \varepsilon_i)}{\sum_{i=1}^{\tilde{N}} H_i^2 + k} = \frac{\sum_{i=1}^{\tilde{N}} H_i^2 \beta_i}{\sum_{i=1}^{\tilde{N}} H_i^2 + k} + \frac{\sum_{i=1}^{\tilde{N}} H_i^T \varepsilon_i}{\sum_{i=1}^{\tilde{N}} H_i^2 + k} \quad (11)$$

$$(1) E(\hat{\beta}^*) = \frac{\sum_{i=1}^{\tilde{N}} H_i^2}{\sum_{i=1}^{\tilde{N}} H_i^2 + k} \beta \neq \beta \quad (12)$$

which denotes that Ridge Regression is biased.

$$(2) V(\hat{\beta}^*) = \frac{\sigma^2 \sum_{i=1}^{\tilde{N}} H_i^2}{(\sum_{i=1}^{\tilde{N}} H_i^2 + k)^2} = \frac{\sigma^2 \sum_{i=1}^{\tilde{N}} \lambda_i}{(\sum_{i=1}^{\tilde{N}} \lambda_i + k)^2} < V(\hat{\beta}) \quad (13)$$

which denotes that Ridge Regression makes the estimation more stable.

$$(3) MSE(\hat{\beta}^*) = \frac{1}{N} \left(\frac{\sigma^2 \sum_{i=1}^{\tilde{N}} H_i^2 + \beta^2 k^2}{(\sum_{i=1}^{\tilde{N}} H_i^2 + k)^2} \right) = \frac{\sigma^2 \sum_{i=1}^{\tilde{N}} \lambda_i}{(\sum_{i=1}^{\tilde{N}} \lambda_i + k)^2} + \frac{\beta^2 k^2}{(\sum_{i=1}^{\tilde{N}} \lambda_i + k)^2} \\ = \frac{1}{N} (\gamma_1(k) + \gamma_2(k)) \quad (14)$$

Hoerl and Kennard had proved that Ridge Regression has less Mean Square Error than the ordinary regression under the proper ridge parameter. It is as follows:

$$\frac{d\gamma_1(k)}{dk} = - \frac{2\sigma^2 \sum_{i=1}^{\tilde{N}} \lambda_i}{(\sum_{i=1}^{\tilde{N}} \lambda_i + k)^3} \quad (15)$$

$$\frac{d\gamma_2(k)}{dk} = \frac{2\beta^2 k \sum_{i=1}^{\tilde{N}} \lambda_i}{(\sum_{i=1}^{\tilde{N}} \lambda_i + k)^3} \quad (16)$$

So, when $k \rightarrow 0^+$

$$\lim_{k \rightarrow 0^+} \frac{d\gamma_1(k)}{dk} = - \frac{2\sigma^2}{(\sum_{i=1}^{\tilde{N}} \lambda_i)^2} < 0 \quad (17)$$

$$\lim_{k \rightarrow 0^+} \frac{d\gamma_2(k)}{dk} = 0 \quad (18)$$

From above equations, $MSE(\hat{\beta}^*)$ is an increasing function of k when $k \in (0, \delta)$, where $\frac{d\gamma_1(k)}{dk} \Big|_{k=\delta} = \frac{d\gamma_2(k)}{dk} \Big|_{k=\delta}$. Therefore, the selection of parameter k is

essential to the performance of Ridge Regression. In our ER-ELM algorithm, a method to determine the ridge parameter proposed by Huang, J.C. [11] is used:

$$k = \hat{\sigma}^2 / \hat{\beta} \quad (19)$$

where $\hat{\beta}$ is the ordinary ELM algorithm estimation and $\hat{\sigma}^2 = \sum_{i=1}^N (y_i - \hat{\beta}x_i)^2 / v, v = n - 1$.

2.3 To Select the Number of Hidden Nodes Based on EM-ELM

The Error Minimized Extreme Learning Machine algorithm starts from a small size of ELM hidden layer, and add random hidden node(nodes) to the hidden layer, while the output weights are updated incrementally.

Suppose a SLFN, $H_1 = H(w_1 \cdots w_{l_0}, b_1 \cdots b_{l_0}, k_1)$ denotes the hidden layer output matrix with l_0 hidden nodes and ridge parameter k_1 calculated by (20). Considering the poor performance due to lower number of hidden nodes, additional l hidden nodes are added to the SLFN. A new hidden layer output matrix H_2 is composed of H_1 and other l extra hidden nodes as

$$H_2 = [H_1, H] \quad (20)$$

where $H = \begin{pmatrix} G(w_{l_0+1}, b_{l_0+1}, x_1, k_2) & \dots & G(w_{l_1}, b_{l_1}, x_1, k_2) \\ \vdots & \ddots & \vdots \\ G(w_{l_0+1}, b_{l_0+1}, x_N, k_2) & \dots & G(w_{l_1}, b_{l_1}, x_N, k_2) \end{pmatrix}_{N \times (l_1 - l_0)}, l_1 - l_0 = l$

and k_2 is ridge parameter of $H^T H$.

Huang had proved $E(H_2) = \min ||H_2\beta_2 - T|| \leq E(H_1) = \min ||H_1\beta_1 - T||$, where $E(H)$ denotes the output error function of SLFNs. We set a stopping criterion for the iterative algorithm as follows:

$$|E(H_{n+1}) - E(H_n)| < \varepsilon \quad (21)$$

where $\varepsilon > 0$ is called the target error.

$$\begin{aligned} H_2^+ &= (H_2^T H_2 + K)^{-1} H_2^T \\ &= (\begin{bmatrix} H_1^T \\ H^T \end{bmatrix} \begin{bmatrix} H_1 & H \end{bmatrix} + \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix})^{-1} \begin{bmatrix} H_1^T \\ H^T \end{bmatrix} \end{aligned} \quad (22)$$

where $K = kI$.

In order to facilitate the calculation, the inversion is substituted as follows:

$$\Phi = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix} = (\begin{bmatrix} H_1^T \\ H^T \end{bmatrix} \begin{bmatrix} H_1 & H \end{bmatrix} + \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix})^{-1} \quad (23)$$

where

$$\varphi_{11} = (H_1^T H_1 + K_1)^{-1} + (H_1^T H_1 + K_1)^{-1} H_1^T H R^{-1} H^T H_1 (H_1^T H_1 + K_1)^{-1}$$

$$\varphi_{12} = -(H_1^T H_1 + K_1)^{-1} H_1^T H R^{-1}$$

$$\varphi_{21} = -R^{-1} H^T H_1 (H_1^T H_1 + K_1)^{-1}$$

$$\varphi_{22} = R^{-1}$$

and

$$\begin{aligned} R &= H^T H + K_2 - H^T H_1 (H_1^T H_1 + K_1)^{-1} H_1^T H \\ &= H^T H + K_2 - H^T H_1 H_1^+ H \end{aligned}$$

so

$$H_2^+ = \begin{bmatrix} U \\ D \end{bmatrix} = \begin{bmatrix} \varphi_{11} H_1^T + \varphi_{12} H^T \\ \varphi_{21} H_1^T + \varphi_{22} H^T \end{bmatrix} \quad (24)$$

and

$$\begin{aligned} D &= R^{-1} H^T - R^{-1} H^T H_1 (H_1^T H_1 + K_1)^{-1} H_1^T \\ &= R^{-1} H^T - R^{-1} H^T H_1 H_1^+ \\ &= (H^T H + K_2 - H^T H_1 H_1^+ H)^{-1} H^T (I - H_1 H_1^+) \\ &= [H^T (I - H_1 H_1^+) H + K_2]^{-1} H^T (I - H_1 H_1^+) \\ &= (H^T M H + K_2)^{-1} H^T M \end{aligned} \quad (25)$$

where $M = I - H_1 H_1^+$

Similarly, we get: $U = H_1^+ - H_1^+ H D$

Now a new hidden layer output matrix is obtained which has less output error. Then we can update the output weight matrix based on the new hidden layer output matrix.

3 Simulation Results by ER-ELM Algorithm

In the simulations, three data sets obtained from website of UCI are applied to evaluate the performance of ER-ELM algorithm. The details of the experimental data sets are listed in Table 1. All data sets are normalized before simulation and all the simulations have been conducted in Matlab 7.8.0(2009a) running on a desktop PC with AMD Athlon(tm) X2 250 processor, 3.00-GHz CPU and 2G RAM.

Table 1. Details of simulation data sets

Data sets	Type	Attributes	Samples	Training data	Testing data
Auto MPG	R	7	398	200	198
Concrete Compressive Strength Data	R	8	1030	600	430
Census	C	14	48842	30000	18842

“R” in the table 1 represents “Regression” while “C” is the abbreviation of “Classification”. In order to obtain more accurate results, every experiment is repeated 50 times. Table 2 shows the simulation results including training time, the number of hidden layers, training mean square error and testing mean square error compared with relative algorithms.

From the experimental results in Table 2, we can see that the ER-ELM algorithm performs better generalization ability and stability compared with the other ELM algorithms.

The ridge parameters play a significant role in the iteration of data processing to deal with the problem of multicollinearity. Table 3 shows the choice of ridge parameter in each iteration based on Auto MPG and Census data sets.

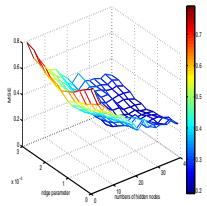


Fig. 1. Average testing accuracy with the variation of k and \tilde{N}

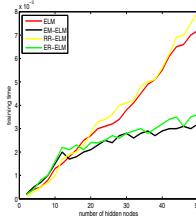


Fig. 2. Average training time of different algorithms

Fig. 1 shows the average testing MSE changes along with the ridge parameter and the number of hidden nodes based on Auto MPG data set. It can be seen that MSE decreases along with the increase of hidden layers under proper ridge parameter. Simultaneously, how to select k also plays an important role in MSE. Fig. 2 shows a training time comparison among the algorithms mentioned in

Table 2. Simulation results

Data Sets	Algorithm type	Stopping criterion	Hidden Nodes	Training Time	Training MSE	Testing Mean	MSE Dev
Auto MPG	ELM	-	30	0.0034	0.1915	0.3015	0.2001
	EM-ELM	0.15	19.8	0.1321	0.1713	0.2121	0.2002
	RR-ELM	-	30	0.0935	0.1615	0.2256	0.2310
	ER-ELM	0.15	20.2	0.8491	0.1289	0.2006	0.2000
Concrete Compressive Strength Data	ELM	-	50	0.0084	0.2388	0.0521	0.0421
Concrete Compressive Strength Data	EM-ELM	0.15	75.5	0.2495	0.2429	0.0432	0.0514
	RR-ELM	-	50	0.0191	0.3004	0.0501	0.0400
	ER-ELM	0.15	80.1	0.8560	0.2359	0.0401	0.0356
Census	ELM	-	80	0.0452	0.0884	0.0412	0.0548
	EM-ELM	0.095	108.5	1.5942	0.0747	0.0402	0.0325
	RR-ELM	-	80	0.2564	0.0754	0.0214	0.0484
	ER-ELM	0.095	120.6	2.5124	0.0200	0.0410	0.0311

Table 3. The choice of ridge parameter in every iteration

Data Sets	Ridge Parameter					
	k_1	k_2	k_3	k_4	k_5	k_6
Auto MPG	0.0025	0.0005	0.0014			
Census	0.0017	0.0009	0.0010	0.0013	0.0000	0.0008

our paper based on Auto MPG data set. We can see that EM-ELM algorithm and ER-ELM algorithm need more time than ELM algorithm and RR-ELM algorithm when choosing less hidden nodes. However, the time spent by ELM algorithm and RR-ELM algorithm rises in a straight line while EM-ELM algorithm and ER-ELM algorithm conducts well in time consuming with increasing the number of hidden nodes.

4 Conclusions

This paper proposes a new ELM algorithm named ER-ELM combined EM-ELM and RR algorithm. ER-ELM algorithm selects the hidden nodes through iterative manner, while deals with the multicollinear problem by adding of ridge parameter k . The theory and simulation results both show that ER-ELM algorithm has better generalization performance and stability than the other ELM algorithms.

References

1. Fu, J., Zhang, H.G., Ma, T.D., Zhang, Q.L.: On passivity analysis for stochastic neural networks with interval time-varying delay. *Neurocomputing* 73, 795–801 (2010)
2. Zhang, H.G., Liu, Z.W., Huang, G.B., Wang, Z.S.: Novel weighting-delay-based stability criteria for recurrent neural networks with time-varying delay. *IEEE Transactions on Neural Networks* 21, 91–106 (2010)
3. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. *Neurocomputing* 70, 489–501 (2006)
4. Bartlett, P.L.: The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* 44, 525–536 (1990)
5. Feng, G.R., Huang, G.B., LiIn, Q.P., Gay, R.: Error minimized extreme learning machine with growth of hidden nodes and incremental learning. *IEEE Transactions on Neural Networks* 20, 1352–1357 (2009)
6. Rong, H.J., Ong, Y.S., Tan, A.H., Zhu, Z.: A fast pruned-extreme learning machine for classification problem. *Neurocomputing* 72(1-3), 359–366 (2008)
7. Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Juttem, C., Lendasse, A.: OP-ELM: Optimally Pruned Extreme Learning Machine. *IEEE Transactions on Neural Networks* 21, 158–162 (2010)
8. Xiao, D., Wang, J.C., Pan, X.L., Mao, Z.Z., Chang, Y.Q.: Modeling and control of guide-disk speed of rotary piercer. *Control Theory & Application* 27, 19–24 (2010)
9. Li, G.Q., Niu, P.F.: An enhanced extreme learning machine based on ridge regression for regression. *Neural Computing & Applications* 22, 803–810 (2013)
10. Uemukai, R.: Small sample properties of a ridge regression estimation when there exist omitted variables 52, 953–969 (2011)
11. Huang, J.C.: Improving the estimation precision for a selected parameter in multiple regression analysis: an algebraic approach. *Economics Letters* 62, 261–264 (1999)

An Attitude Determination System of Quad-rotor Aircraft Based on Extended Kalman Filter and Data Fusion Technique

Xinfu Ye¹, Lian Zhu¹, Sun Changyin¹, and Wei Wang²

¹ School of Automation, Southeast University, Nanjing 210096, China

² Nanjing University of Information Science & Technology, Nanjing 210044, China

yexinfu88@163.com

Abstract. An attitude determination system composed of gyroscopes, accelerometer and magnetometer is designed. An extended Kalman filter was used for estimating aircraft's state. The attitude can be determined via data fusion technique. The experimental results indicate that this combination could effectively restrain attitude error arising from the random drift of sensors. The system has the characteristics of small size, low cost and good reliability.

Keywords: attitude, extended Kalman filter, data fusion.

1 Introduction

UAVs have aroused a great interest in the research and academic circles in recent years, for it could be widely used in many important places. Such as monitoring, vigilance, inspection, rescue, electronic intelligence, communication relays [1-3]. A necessary part of these missions is accurate navigation of the vehicle. When we control a UAV need some knowledge of its attitude angles [4-6]. Modern MEMS technologies are offering light and moderate-cost solutions, denoted as inertial measurement units (IMUs).

Now many researchers have been worked on the attitude estimation technology, there are several integrated system have been developed and implemented. Hector Garcia de Marina et al proposed an attitude estimation method using unscented Kalman filter and three-axis attitude determination [3], Yong and Elkaim made combination of GPS, magnetometers and MEMS inertial sensors to measure the attitude [7-8]. Li Wenqiang and MA Fuchang, et al design an attitude indicator based on MEMS technology. It analyzes the principle of testing attitude by acceleration sensor, the conditioning circuit, calculation and scale transform. It also gives a method which can realize by MMA 7260 to measure the biaxial acceleration signal [9], but the accurate of the system is not very high.

Therefore, this paper will present an integrated attitude estimation technology based on MEMS gyroscopes, accelerometer, and magnetometer. Accelerometer was used to judge the motion state of the craft; the gyroscopic drift was compensated by the attitude information obtained through combined calculation of the accelerometer and magnetometer in the filtering algorithm. An Extended Kalman Filter algorithm

(EKF) will be designed to estimating aircraft's state, data fusion algorithm was used to prove the accuracy of attitude.

2 Attitude Measurement Hardware System

Attitude measurement system consists of three Single-axis gyroscope tri-axial accelerometer geomagnetic sensor module and microcontroller. MAG3110module is selected as geomagnetic sensor to get the aircraft heading. ADXL335 is a small size, low power, three-axis accelerometer. The output signals of the accelerometers $[a_x \ a_y \ a_z]$ and the gyroscopes $[w_x \ w_y \ w_z]$, which are the body frame accelerations and the roll, pitch, yaw angular rates, respectively, are measured directly with an STM32 A/D converter microcontroller. Fig.1 shows the attitude measurement system block diagram, Fig.2 shows the prototype of the attitude measurement module.

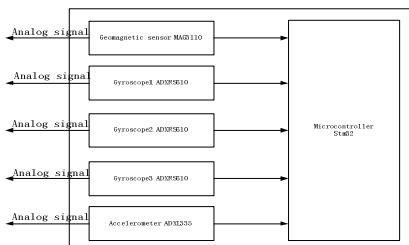


Fig. 1. Attitude measurement system block diagram



Fig. 2. Attitude measurement module

3 Attitude Determination System

Fig.3 shows the attitude determination system structure. The gyroscopes will measure the angular rate of the body coordinate system relative to the reference system. The tri-axial accelerometers will measure the acceleration of the body coordinate system. A tri-axial geomagnetic will measure the magnetic field strength in the X, Y, Z direction. The equivalent vector method [10] is adopted to calculate the attitude, extended Kalman filter combines the attitude of gyroscopes with those of accelerometers in order to estimate the attitude errors. The attitude can be determined through data fusion at last.

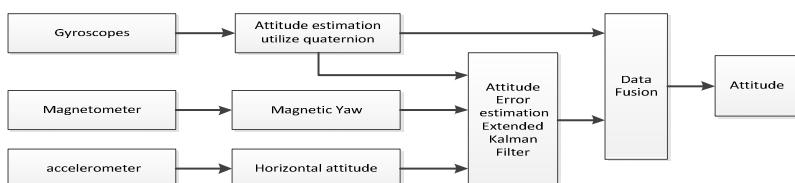


Fig. 3. Attitude determination system structure

3.1 Attitude Algorithm Based on Gyroscopes

Assume the vector of reference system is u and the vector of body coordinate system is v , so the relationship between two coordinate system can be expressed as $u = R_b^r v$.

Where R_b^r is the direction of cosine matrix:

$$R_b^r = \begin{bmatrix} \cos \psi \cos \theta & \sin \psi \cos \theta & -\sin \theta \\ \cos \psi \sin \theta \sin \phi - \sin \psi \cos \phi & \sin \psi \sin \theta \sin \phi + \cos \psi \cos \phi & \cos \theta \sin \phi \\ \cos \psi \sin \theta \cos \phi + \sin \psi \cos \phi & \sin \psi \sin \theta \cos \phi - \sin \phi \cos \psi & \cos \theta \cos \phi \end{bmatrix} \quad (1)$$

Where θ is the pitch angle, ϕ is the roll angle, ψ is the yaw angle. Rotational transformation in relations between the two coordinate systems can also use the quaternion as follows:

$$R_b^r = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1 q_2 + q_0 q_3) & 2(q_1 q_3 - q_0 q_2) \\ 2(q_1 q_2 - q_0 q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2 q_3 - q_0 q_1) \\ 2(q_1 q_3 + q_0 q_2) & 2(q_2 q_3 - q_0 q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \quad (2)$$

So the relationship between attitude angle and quaternion can be expressed as follow:

$$\begin{aligned} \theta &= \arcsin(2(q_1 q_3 - q_0 q_2)) \\ \phi &= \arctan\left(\frac{2(q_2 q_3 + q_0 q_1)}{q_0^2 - q_1^2 - q_2^2 + q_3^2}\right) \\ \psi &= \arctan\left(\frac{2(q_1 q_2 + q_0 q_3)}{q_0^2 + q_1^2 - q_2^2 - q_3^2}\right) \end{aligned} \quad (3)$$

In this paper, the equivalent vector method Miller is adopted to determine the attitude based on gyroscopes. The calculation procedure is as follows:

- 1). Calculated the increment of angle $\theta_1, \theta_2, \theta_3$ at $t = \frac{1}{3}h, t = \frac{2}{3}h, t = h$ respectively.
 - 2). Calculated angle ϕ according to equation $\phi = \theta + 0.45 * (\theta_1 - \theta_3) + 0.675 * \theta_2 - (\theta_3 - \theta_1)$
 - 3). Calculated quaternion according to equation
- $$q(t) = \begin{bmatrix} \cos\left(\frac{\phi_0}{2}\right) & \frac{1}{\phi_0} \sin\left(\frac{\phi_0}{2}\right) \cdot \phi_x & \frac{1}{\phi_0} \sin\left(\frac{\phi_0}{2}\right) \cdot \phi_y & \frac{1}{\phi_0} \sin\left(\frac{\phi_0}{2}\right) \cdot \phi_z \end{bmatrix}, \quad \phi_0 = \sqrt{\phi_x^2 + \phi_y^2 + \phi_z^2}$$
- 4). Update the quaternion $Q(T+t)$ utilize equation $Q(T+t) = Q(T) \otimes q(t)$;
 - 5). Make the $Q(T) = Q(T+t)$,
 - 6). Calculated corresponding attitude in line with equation (4).

3.2 Attitude Algorithm Based on Accelerometer and Magnetometer

The accelerometer measurement output is $A_r = [0 \ 0 \ g]$ in reference coordinate system when the aircraft is in horizontal static. The accelerometer measurement

output is $A_b = [a_x \ a_y \ a_z]$ when the aircraft in any posture. According to equation $A_b = R_r^b A_r$, θ and ϕ can be written as:

$$\theta = \arcsin a_x \quad (4)$$

$$\phi = \arcsin\left(\frac{a_y}{g \cos \theta}\right) \quad (5)$$

The yaw can be easily calculated by combine the output of magnetometer and θ , ϕ .

$$\psi = \frac{m_y \cos \phi + m_z \sin \phi}{m_x \cos \theta + m_y \sin \theta \sin \phi - m_z \cos \theta \sin \phi} \quad (6)$$

Where, m_x , m_y , m_z are respectively measured value of magnetometer X-axis, Y-axis, Z-axis.

3.3 Attitude Error Estimation Based on Extended Kalman Filter

From the reference coordinate system to body coordinate system, the quaternion is defined as $q = q_0 + q_1 \vec{i} + q_2 \vec{j} + q_3 \vec{k} = q_0 + \vec{q}$. The quaternion kinematics equation is given as follows:

$$\dot{q} = \frac{1}{2} \Omega(w) q \quad (7)$$

Where, $w = [w_x \ w_y \ w_z]^T$ is the angular rate vector of body coordinate system, and

$$\Omega(w) = \begin{bmatrix} 0 & -w_x & -w_y & -w_z \\ w_x & 0 & w_z & -w_y \\ w_y & -w_z & 0 & w_x \\ w_z & w_y & -w_x & 0 \end{bmatrix} \quad (8)$$

Deal with the attitude quaternion so that build the Kalman filter equation. Define error between the real quaternion and the estimation is $Q_e \approx (1, \vec{q}_e)^T$, and $\vec{q}_e = [q_{e1} \ q_{e2} \ q_{e3}]^T$. The random drift of gyroscopes as $\delta w = [\delta w_x \ \delta w_y \ \delta w_z]^T$. Considering the aircraft rotation is a small angle relative to the reference coordinate system of each sampling time, so the q_e is a small error. $R_b^r(q_e)$ can be linearized as

$$R_b^r(q_e) = \begin{bmatrix} 1 & 2q_{e3} & -2q_{e2} \\ -2q_{e3} & 1 & 2q_{e1} \\ 2q_{e2} & -2q_{e1} & 1 \end{bmatrix} = I + 2[\vec{q}_e \cdot] \quad (9)$$

the direction of cosine matrix can be expressed:

$$R_b^r = R_b^{r'} \cdot R_b^{b'} = R_b^{r'} \cdot R_b^r(q_e) \quad (10)$$

Where b is the body coordinate system, b' is the body coordinate system of previous sampling point. According to equation (10), (11) we can obtain:

$$\begin{aligned} R_b^r &= R_{b'}^r(I + 2[\vec{q}_e \cdot]) \\ R_r^b &= (I - 2[\vec{q}_e \cdot])R_r^{b'} \end{aligned} \quad (11)$$

Define \vec{m}_b , \vec{m}_r respectively as measured value geomagnetic vector in the body coordinate system and reference coordinate system.

$$\vec{m}_b = R_r^b \cdot \vec{m}_r = (I - 2[\vec{q}_e \cdot])R_r^{b'} \cdot \vec{m}_r = (I - 2[\vec{q}_e \cdot]) \cdot \vec{m}_r \quad (12)$$

The measured error in body coordinate system is:

$$\delta \vec{m}_b = \vec{m}_b - \vec{m}_{b'} = -2[\vec{q}_e \cdot] \cdot \vec{m}_{b'} = 2[\vec{m}_{b'} \cdot] \cdot \vec{q}_e \quad (13)$$

Similarly, the gravity vector measured error in body coordinate system is

$$\delta \vec{a}_b = 2[\vec{a}_{b'} \cdot] \cdot \vec{q}_e \quad (14)$$

Taken \vec{q}_e and δw as the state variables, $\delta \vec{m}_b$ and $\delta \vec{a}_b$ as observed variables, that is $x = [\vec{q}_e, \delta w]^T$. $y = [\delta \vec{m}_b \quad \delta \vec{a}_b]$ Therefore, the state equation and observed equation can be written in the form:

$$\begin{aligned} \dot{x} &= F \cdot x + G w \\ y &= H \cdot x + v \end{aligned} \quad (15)$$

$$\text{where } F = \begin{bmatrix} \Omega' & R_b^r(\vec{q}_e) \\ 0 & 0 \end{bmatrix}, \quad G = I_{6 \times 6}, \quad \Omega' = \begin{bmatrix} 0 & w_z - \delta w_z & -(w_y - \delta w_y) \\ -(w_z - \delta w_z) & 0 & w_x - \delta w_x \\ -(w_y - \delta w_y) & -(w_x - \delta w_x) & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 2[\vec{m}_b \cdot] & 0 \\ 2[\vec{a}_b \cdot] & 0 \end{bmatrix}$$

w, v is the Process noise. Thus the discrete state equation is:

$$\begin{aligned} x_k &= F_{k,k-1} \cdot x_{k-1} + G_{k,k-1} \cdot w_{k-1} \\ y_k &= H_{k,k-1} \cdot x_k + v_k \end{aligned} \quad (16)$$

Attitude error can be estimation as follows:

- a). set initial state x_0 , error estimation covariance P_0 , process noise covariance Q , observation noise covariance R .
- b). calculate the state variable: $\hat{x}_{k,k-1} = F_{k,k-1} \hat{x}_{k-1}$
- c). calculate the error covariance: $P_{k,k-1} = F_{k,k-1} P_{k-1} F_{k,k-1}^T + G_{k,k-1} Q_{k-1} G_{k,k-1}^T$
- d). calculate the gain: $K_k = P_{k,k-1} H_k^T (H_k P_{k,k-1} H_k^T + R_k)^{-1}$
- e). update the error covariance: $P_k = (I - K_k H_k) P_{k,k-1}$
- f). state estimation: $\hat{x}_k = \hat{x}_{k,k-1} + K_k (y_k - H_k \hat{x}_{k,k-1})$

3.4 Attitude Determined by Data Fusion

The attitude obtained by gyroscopes accelerometer and magnetometer both are not accurate. So we fusion the attitude which obtained from equivalent rotating vector and obtained from accelerometer and magnetometer combination method. We use different method to correct the attitude when aircraft is in different motion condition. The aircraft motion can be judged by accelerometer, whether aircraft satisfies the equation $|\sqrt{a_x^2 + a_y^2 + a_z^2} - g| < \delta$ within Δt seconds, $\delta > 0$ depends on noise of accelerometer. When aircraft is in the variable motion state, we obtain the attitude by equivalent rotating vector. In other hand, aircraft is stationary or in uniform motion, we design an extended Kalman filter to correct the measured drift of gyroscopes. The gravity vector and geomagnetic vector measured error which are utilized as observed variables, while the attitude error and drift of gyroscope used as state variables. The calculation steps are summarized as follows:

- 1): Judge the aircraft motion state by using the equation $|\sqrt{a_x^2 + a_y^2 + a_z^2} - g| < \delta$
- 2): calculate the attitude of aircraft by using equivalent rotating vector.
- 3): if aircraft is in the variable motion state set the attitude error is zero. Otherwise, correct the attitude error through extended Kalman Filter, accelerometers and magnetometers introduced in the previous section.
- 4): get the attitude estimation through updating the attitude via the attitude error calculated by the previous step,
- 5): convert the attitude estimation to quaternion estimation \hat{Q} , set the quaternion $Q(t) = \hat{Q}$.

4 Experiment Results

To verify the feasibility of the hardware circuit and the attitude algorithm, some experiments were performed in the fight test. The data of attitude before calibration compared with the data after fusion shown in the Fig. 4. At the same time, the actual attitude control results show in Fig. 5.

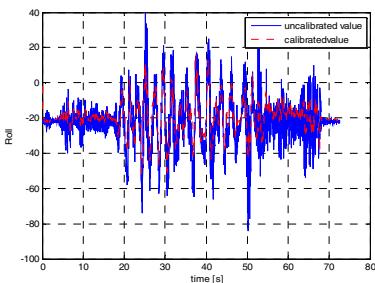


Fig. 4. The attitude angle before and after calibrated

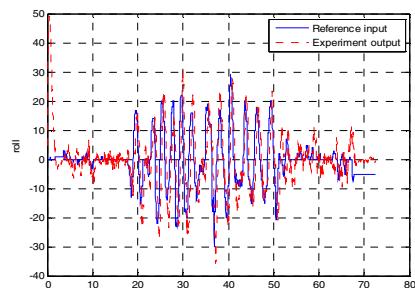


Fig. 5. Attitude control experimental results

5 Conclusions

In this paper, attitude determination method is given through utilizing accelerometer and magnetometer compensation gyro drift. The basic ideal of improving attitude accuracy lies in the extended Kalman filter with ability to estimate the attitude errors and gyro bias states efficiently through multi-sensor data fusion. But there are still too many uncertain factors in attitude measured system. Therefore, We will consider combining other sensors, such as video, laser ranger and develop the improved fusing algorithm in future work.

References

1. Ivan, G.: Attitude stabilization of a Quad-rotor UAV based on rotor speed sensing with Accelerometer data estimation via Kalman filtering. In: Proc. of the 31st Chinese Control Conference, pp. 5123–5128 (2012)
2. Maza, I., Caballero, F., Capitan, J., de Dios, J.M., Ollero, A.: Firemen monitoring with multiple UAVs for search and rescue missions. In: IEEE Workshop SSRR, pp. 1–6 (2010)
3. de Marina, H.G., Pereda, F.J., Giron-Sierra, J.M.: UAV Attitude Estimation Using Unscented Kalman Filter and TRIAD. Proc. IEEE Trans. Ind. Electron. 59(11), 4465–4473 (2012)
4. Xia, Y., Zhu, Z., Fu, M., Wang, S.: Attitude tracking of rigid spacecraft with boundeddisturbances. IEEE Trans. Ind. Electron. 58(2), 647–659 (2011)
5. Zheng, B., Zhong, Y.: Robust attitude regulation of a 3-dof helicopter benchmark: Theory and experiments. Proc. IEEE Trans. Ind. Electron. 58(2), 660–670 (2011)
6. Zhu, Z., Xia, Y., Fu, M.: Adaptive sliding mode control for attitude stabilization with actuator saturation. Proc. IEEE Trans. Ind. Electron. 58(10), 4898–4907 (2011)
7. Yong, L., Dempsteri, A., Binghao, L., et al.: A low-cost attitude heading reference system by combination of GPS and magnetometers and MEMS inertial sensors for mobile applications. Journal of Global Positioning Systems 5, 88–95 (2006)
8. Elkaim, G.H., Foster, C.: MetaSensor: Development of a Low-Cost, High Quality Attitude Heading Reference System. Institute of Navigation ION GNSS (2006)
9. Li, W., Ma, F., et al.: A Design of Attitude Indicator Based on MEMS Technology. In: Proc. International Conference on Computer Application and System Modeling, pp. 338–341 (2010)
10. Miller, R.B.: A new strap down attitudealgorithim. Journal of Guidance 6(4), 289–291 (1983)

Connectivity of Clustered and Multi-type User CR Network: A Percolation Based Approach

Jingyuan Guo*, Tao Yang, Hui Feng, and Bo Hu

Electronic Engineering Department,
Fudan University,
200433 Shanghai, China

Abstract. For the clustered, large scale ad hoc cognitive radio network with multi-type users, we address the percolation-based connectivity problem, in which the existence of a communication link between two secondary users depends on not only the distance between them, the transmitting and receiving activities of nearby primary users, but also the neighboring user's type. From the mean-field approximation perspective, we firstly give the sufficient condition for the single type (here "type" means transmission radius) clustered secondary users on how the marginal nodes in the clusters are correlated to provide the critical percolation radius. Then the connectivity of the secondary users with inhomogeneous node distribution is studied, where two types of sub-critical secondary users are migrated into a super-critical cognitive user network through a multi-type branching process in random environment, which essentially is a percolation parameter optimization problem. Various simulations are performed to show the percolation is effective both in theory and practice in guidance of the deployment of the wireless network.

Keywords: Percolation, connectivity, cognitive radio, heterogeneous networks, mean field approximation, branching process.

1 Introduction

With the increasing demand on the flexible and instant access in various network, the interaction between various kinds of terminal users will get inevitable and frequent, which makes the underlying network present more and more heterogeneous and dynamic characteristics. The cognitive radio (CR) technology is a typical example in characterizing the interaction between primary user (PU) and secondary user (SU) through opportunistic spectrum sharing. In the large-scale CR network, we focus on the condition that SU network can be globally connected. In literature, the continuum percolation theory[1] has emerged in recent years as a powerful tool in modeling the various connectivity behavior. For the homogeneous ad hoc network consisting peer users, the connectivity has been well studied [2] - [6], while for the heterogeneous network, only some scattered studies appeared, and it's far from perfect, among which Wei Ren etc. [7]

* This research was supported by the 973 project of China under Grant 2011CB302903.

analyzed the connectivity region, and obtained the quantitative sufficient and necessary condition on connectivity via the parameters of node density, interference range as well as transmission range. It's the first work to couple the occupancy of or withdrawing from opportunistic spectrum with the connected component and vacant component in Boolean percolation model. For the impact of cooperation between secondary users on the connectivity, Weng Chon etc [8]. modeled the heterogeneous network as the multiple ad hoc networks and multiple infrastructure networks, where mean degree and the degree distribution of a network node in both noise-limited and interference-limited environment with general fading channel is derived, based on which the necessary condition for cooperative percolation is analyzed. Differing from [7][8], the cognitive radio graph model (CRGM) is introduced in [9] to formulate the percolation problem in the graph, where the number of active channels and the activities of primary users is taken into account, and thus an upper bound of the critical density of the primary users is achieved given that the secondary user can form a percolated network. The other important aspect concerns how the link dynamics affect the connectivity, Pu Wang etc. [10] has proved that if the secondary density is greater than a critical threshold, then the connectivity can be maintained at all times even under the dynamically changing radio environment. Furthermore, it is shown that even when the secondary network is disconnected at all times, it is still possible for a SU to transfer its message to any destination with a certain delay with probability one, moreover, this delay is asymptotically linear in the Euclidean distance between the transmitter and receiver.

1.1 Our Main Idea

As indicated in literature, on one hand, although Poisson point process(PPP) has been widely adopted as the tool in analyzing the stochastic network connectivity, it is not always the case, as cluster distribution has been proven to be much closer to the real cognitive scenario such as the urban hotspot [11][12] even though its connectivity is still unknown. On other hand, conventional analysis only focus on the interaction between two kinds of heterogeneous user, i.e., primary user and secondary user, actually the secondary user is often comprised of various terminal users, e.g. two kinds of user with different transmission range. When this two aspects are combined together, we concentrate on how the connectivity behaviors change with the network parameters. Amongst these issues, our contributions are as follows: For Poisson distributed secondary users with two different transmission radius, percolation condition is given. From the mean field approximation perspective, we analyze the expected percolation radius, and multi-type branching process is introduced to model the connectivity with inhomogeneous node distribution.

2 Percolation Model for Both Two-Type User and Clustered Node Distribution

Consider a special case of multi-type secondary user network – type 1 and type 2 user, both are omnidirectional radiation and with different transmission range r_1 and r_2 ($r_1 > r_2$), respectively. The secondary network network is overlaid with the primary user network, which is formed by a PPP, the node density and transmission radius is determined but unknown. Based on these premises, we consider the connectivity of the secondary network in the following several scenarios:

- Same distribution (PPP) but with two transmission radius (r_1, r_2) for type 1 and type 2 users, see Section 2.1;
- Cluster distribution (Thomas Process) with only one transmission radius (r), see Section 2.2;

The general heterogeneous network connectivity behavior can be almost completely characterized by the three cases above. When there are multiple types of the secondary users, an extension can be easily reached to give a similar result.

2.1 Two-Type Secondary User Network with Poisson Distribution

Assume the given primary network consists of transmitters of Poisson distribution and transmitter node density is λ_{PT} . Here if not specified, the transmitter will indicate a random transmitter-receiver pair. The two-type secondary users are generated through PPP with transmission range r_1 and r_2 , respectively. As for the formulation of the bidirectional link, we adopt the model in Fig.3 of [7] as a reference, so we have the following proposition.

Proposition 1: Assume the proportion of two types of secondary users in the percolated secondary network is p_1 and p_2 ($p_1 + p_2 = 1$), given transmission radius r_1 and r_2 ($r_1 > r_2$), respectively. If $p_1 > \frac{2r_2}{r_1+r_2}$, then the primary node density should satisfy $\lambda_{PT} \leq \frac{1}{4r_1^2 - r_1^2} \lambda_C(1)$, else $\lambda_{PT} \leq \frac{1}{4r_2^2 - r_2^2} \lambda_C(1)$. Here r_i and $\lambda_c(1)$ indicate the interference range of the secondary user and the critical density corresponds to radius of 1 in the Boolean model, respectively.

Proof: According to the scaling theorem of the Poisson Boolean models [1], in the two dimension plane, the critical density $\lambda_C(r_i)$ for the given transmission radius meet the following equation

$$\lambda_C(r_1)r_1^2 = \lambda_C(r_2)r_2^2 \quad (1)$$

when two types of user node are mixed, the critical density λ_C should satisfy

$$\lambda_C(r_1) < \lambda_C < \lambda_C(r_2) \quad (2)$$

and varies with the ratio of $p_1(r_1)/p_2(r_2)$. Assume O_i, a_i are the transmitter/receiver of type 1 and type 2 user, respectively (See Fig. 1). the percolation analysis through the node O_i and a_j can be divided into two typical cases as follows.

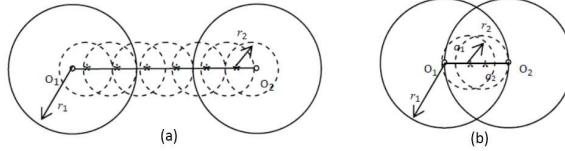


Fig. 1. O_1 and O_2 are type 1 users with communication range r_1 ; a_1 and a_2 are type 2 users with communication range r_2 , $r_1 > r_2$

- **Case 1:** If O_1 and O_2 can't communicate with each other (See (a) in Fig. 1), which means O_1 and O_2 does not lie within each others transmission range circle, in order to enable the link between O_1 and O_2 , there must be type 2 users working as relays in between, due to $r_1 > r_2$, obviously

$$\lambda_c = \lambda_2 \quad (3)$$

which means if percolation occurs, it must occur firstly in type 2 user and $p_1(r_1)/p_2(r_2) < 1$. Otherwise, if percolation occurs in type 1 user, given that both two types of users are PPP, obviously the percolation in type 2 user is covered by the percolation of type 1 user, and this is another typical case 2 analyzed as follows.

- **Case 2:** Consider $|O_1O_2| \leq r_1$, secondary user nodes $a_i (i = 1, 2, \dots)$, acting as a relay node, existed between O_1 and O_2 (See (b) in Fig. 1). If O_1 and O_2 can form a bidirectional link, then all the nodes a_i in between will form a bidirectional link accordingly. So the critical density of type user 2 depends only on $|O_1O_2|$, in terms of *finite size scaling*[1], if N nodes exist between O_1 and O_2 , when percolation occurs, we have $p_1 : p_2 = 2 : N$, then $(N+1)p_1 + p_1 = 2$, if the critical density of a_i is greater than or equal to $\lambda_C(r_2)$, obviously the percolation occurs on type user 2, this is a trivial case just like case 1. Consider another scenario, with $|O_1O_2| \leq r_1$, if $\frac{r_1}{N+1} > r_2$, which means if percolation take place, then the critical density is a function of both r_1 and r_2 , i.e., $\lambda_C(r_1) < \lambda_C < \lambda_C(r_2)$, so we can get the share of the type 1 user proportion as

$$p_1 > \frac{2r_1}{r_1 + r_2} \quad (4)$$

To combine both case 1 and case 2 together, we concluded that if (4) is satisfied, then the critical density $\lambda_C(r_1, r_2)$ is a function of the two-type user radius. If the Poisson primary user is overlaid on the secondary user, according to the Theorem 2.2 [7], the interaction between primary and secondary user will firstly exist in type 1 user, so the condition on density of primary user node is achieved. Otherwise, this interaction will take place on type 2 user, which corresponds to the inequality in Proposition 1, and thus the proof is completed. ■

2.2 Cluster Distribution (Thomas Process) with Single Type User

When the secondary user with single transmission radius is distributed in a cluster pattern, we focus on how the *critical parameter* varies with the cluster parameter. Here for convenience, we adopt the Thomas point process as the underlying cluster model, that is, the cluster center follows a Poisson distribution with intensity λ , and the points around the center follow a Gaussian distribution with mean value μ and covariance matrix $\Sigma = \text{diag}(\sigma_x^2, \sigma_y^2)$. All points acting as a transmitter/receiver node with the same transmission range of r . A realization of Thomas process in unit square is depicted in (a) of Fig. 2.

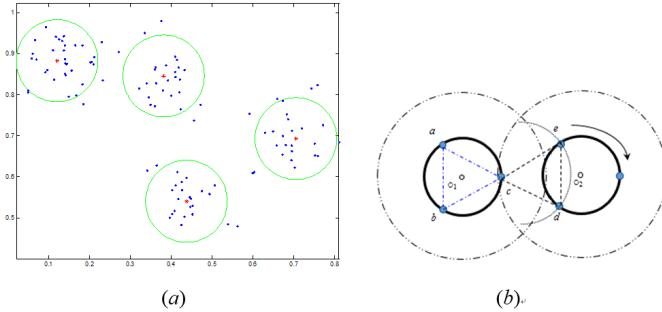


Fig. 2. (a) A realization of a Thomas process in unit square with $\lambda = 4$, $\mu = 24$ and $\sigma_x = \sigma_y = 0.05$. (b) The mean field approximation approach on the connectivity.

Intuitively, when the node distribution varies from Poisson to Thomas distribution, the occurrence of percolation will get more and more difficult. The reason lies in that with the increasing clustering, the function of the connectivity for part of nodes will be offset, and thus the overall connectivity performance will depend on the marginal nodes far away from the centre node. Once this node can connect those marginal nodes distributed around the neighboring cluster, the two clusters will be fully connected. With this idea in mind, we adopt the theory of Mean Field Approximation (MFA) to get the equivalent critical radius, see (b) of Fig. 2.

Proposition 2: The sufficient condition of the expected critical radius on connectivity in the Thomas model satisfy

$$E[R] = \left(\frac{r^2}{r'} N! \right) \cdot \sum_{k=\frac{1-(-1)^N}{2}}^{\frac{N+1-(-1)^N}{2}} \frac{\pi^{2k-\frac{1-(-1)^N}{2}}}{\left[2k - \frac{1-(-1)^N}{2} \right]} \cdot (-1)^m \quad (5)$$

where

$$m = k + \frac{N + \frac{1-(-1)^N}{2}}{2} + (-1)^{\frac{N}{2}} \times \left[\frac{1 - (-1)^{N+1}}{2} \right]$$

$N = 2, 3, \dots$

where r and r' indicate the radii, N is the number of nodes uniformly distributed in the equivalent circle.

Proof: Omitted due to the length limited. ■

Here if the primary user is overlaid over the above Thomas model, a similar analysis can be performed as in case 1.

We use Monte Carlo simulation to verify the critical connection performance for a two-dimension Thomas process, which aims at finding where and when the percolation will happen, and how the interaction between parameters r , λ , μ and Σ leading to the occurrence of phase transition. Through finite size scaling, we first set the initial value of the parameters and spread the nodes in a unit square according to Thomas distribution, by keeping any three parameters fixed and let another one parameter varied, we get the observation for percentage of nodes in the largest cluster, with respect to different variable parameters such as r , μ , λ and Σ .

Fig. 3 shows the effect on connectivity by the communication range r , here $\sigma_x^2 = \sigma_y^2 = 0.2$. It is clear that if full connectivity is not required one can significantly decrease the radius r . The clustering parameter Σ has a minor effect on the growth of the giant component, as smaller Σ accelerates the emergence of the giant component. Due to the length of the paper, we omitted the effect on connectivity by the parameter μ , λ , Σ .

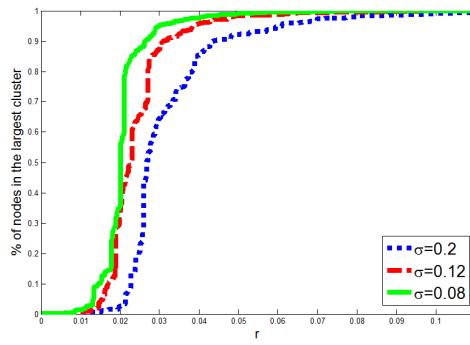


Fig. 3. 2-D Thomas process. Percentage of nodes in the largest cluster, with $\sigma_x = \sigma_y = 0.2$.

3 Multi-type User Connectivity under Inhomogeneous Node Distribution

In a practical secondary ad hoc network, the user nodes are presented with either Poisson or Thomas distribution seems a little far-fetched, generally the inhomogeneous Poisson distribution is employed to characterize this kind of distribution, but the connectivity problem isn't still fully solved [7]. Here we construct the

inhomogeneous connectivity as a multi-type branching process in random environment (MBPR) problem, with which we analyze theoretically how connectivity performance change with distribution parameter.

In the underlying problem, random environment indicates the reproduction distribution is environment dependent, the type, as previously defined, is the transmission radius, and thus the MBRP is formulated as in what condition these inhomogeneous nodes with different radius can be connected through a cooperation strategy.

Proposition 3: Taking the primary users into consideration, its interference range is R_I . In the case where there are two kinds of secondary users type 1 and type 2, with the mixed density of λ_S , i.e. density of type 1 SUs is $p\lambda_s$ and type 2 is $(1-p)\lambda_s$, with transmission range r_{p1} , r_{p2} , interference range r_{I_1} , r_{I_2} respectively. Then the necessary condition for percolation of this mixed SU network is

$$\max\{|eig(\mathbf{A})|\} > 1 \quad (6)$$

where A is the the first moment matrix $\mathbf{A} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$ and

$$\begin{aligned} m_{12} = & 2(1-p)\lambda_s \int_0^{r_{p1}} t \exp\{-\lambda_{PT}[\pi(r_{I_2}^2 + R_I \\ & + I(R_I, R_p, r_{I_1})) - S_I(t, r_{I_1}, r_{I_1}) - S_I(t, R_I, R_I) \\ & - \frac{1}{\pi R_p^2} \iint_{S_{U_2}(t, R_I, R_I)} S_{I_2}(r, \theta, R_p, t, r_{I_1}, r_{I_2}) r dr d\theta]\} dt \end{aligned} \quad (7)$$

where

$$I(R_I, R_p, r_{I_1}) = \frac{2}{\pi R_p^2} \int_0^{R_I} t S_I(t, R_p, r_{I_1}) dt$$

$S_I(t, r_1, r_2)$ is the common area of two circles with radii r_1 and r_2 and centered t apart, $S_{U_2}(t, r_1, r_2)$ the union area of two circles with radii r_1 and r_2 and centered t apart. m_{11} , m_{21} and m_{22} can be found in the similar way.

Proof: The basic idea is to construct a MBPR process, where each individual is assigned a type in a set $T = \{1, 2\}$ and may give birth to individuals of their own type and also of other types. e.g. a type 1 SU can communicate with type 1 or type 2 SU that satisfies the following two condition:

C1: The distance between them is at most $\max\{r_{p1}, r_{p2}\}$;

C2: There exists a bidirectional spectrum opportunity between them.

For each individual, say of type $r \in T$, is associated with a random vector $\xi_r = \{\xi_r^1, \xi_r^2\}$, where ξ_r^j is a random variable that represents the number of children of type j born from a type r individual. Writing $m_r^j = E[\xi_r^j]$, $r, j \in T$, the first moment matrix can be constructed as in (7). According to [13], a MBPR is classified as *critical*, *subcritical* or *supercritical* if $\rho = 1$, $\rho < 1$ or $\rho > 1$, respectively, where ρ is the spectral radius of matrix \mathbf{A} (i.e. the supremum among the

positive eigenvalue of \mathbf{A}). In this 2-branched case, the four elements of \mathbf{A} can be found in a similar way as in [7]. ■

4 Conclusions

In this paper, we studied the critical connectivity of ad hoc heterogeneous communication networks under the cognitive radio framework, for the typical three scenarios with and without the overlaid primary user network, we analyze the condition on connectivity through both theory analysis and simulation.

Mean field approximation is employed to give the expected connection radius given that the single type user are modeled as a Thomas distribution. For the inhomogeneous distributed nodes, we introduce the MBPR model to characterize the necessary condition for percolation.

References

1. Meester, R., Roy, R.: Continuum percolation. Cambridge University (1996)
2. Cheng, Y.-C., Robertazzi, T.G.: Critical connectivity phenomena in multihop radio models. *IEEE Trans. on Communications* 37 (July 1989)
3. Dousse, O., Baccelli, F., Thiran, P.: Impact of interferences on connectivity of ad hoc networks. *ACM/IEEE Trans. Netw.* 13(2), 425–436
4. Dousse, O., Franceschetti, M., Macris, N., Meester, R., Thiran, P.: Percolation in the signal to interference ratio graph. *J. Appl. Prob.* 43, 552–562 (2006)
5. Kong, Z., Yeh, E.: Connectivity, percolation, and information dissemination in large-scale wireless networks with dynamic links. *IEEE Transactions on Information Theory* (2009)
6. Vaze, R.: Percolation and connectivity on the signal to interference ratio graph. In: Proceedings of the IEEE Conference on Computer Communications (INFOCOM 2012), Orlando, Florida, USA (March 2012)
7. Ren, W., Zhao, Q., Swami, A.: Connectivity of heterogeneous wireless networks. *IEEE Trans. Inf. Theory* 57(7), 4315–4332 (2011)
8. Ao, W.C., Chen, K.-C.: Cognitive Radio-Enabled Network-Based Cooperation: From a Connectivity Perspective. *IEEE Journal on Selected Areas in Communications* 30(10), 1969–1982 (2012)
9. Lu, D., Huang, X., Li, P., Fan, J.: Connectivity of large-scale cognitive radio ad hoc networks. In: IEEE INFOCOM (2012)
10. Wang, P., Akyildiz, I.F., Al-Dhelaan, A.M.: Percolation Theory based Connectivity and Latency Analysis of Cognitive Radio Ad Hoc Networks. *Wireless Networks* 17, 659–669 (2011)
11. Li, D., Gross, J.: Robust clustering of ad-hoc cognitive radio networks under opportunistic spectrum access. In: Proc. IEEE Int. Conf. Commun., Kyoto, Japan (June 2011)
12. Liu, S.S., Lazos, L., Krunz, M.: Cluster-Based Control Channel Allocation in Opportunistic Cognitive Radio Networks. *IEEE Trans. Mobile Computing* 11(10), 1436–1449 (2012)
13. Harris, T.E.: The Theory of Branching Processes. Courier Dover Publications (2002)

Patch-Based Tracking and Detecting for Visual Tracking

Qianwen Li and Yue Zhou

Institute of Image Processing and Pattern Recognition

Shanghai Jiao Tong University

Shanghai, China

liqianwen115@163.com, zhouyue@sjtu.edu.cn

Abstract. As one of the most traditional tracking methods, particle filter has been improved in many previous tracking methods due to its non-Gaussian and non-linear distribution. Meanwhile, pure tracking methods cannot achieve good performance in complex tracking scenarios where there enormous deformation and occlusion occur. We present a combination of patch-based tracking and detecting methodology for visual tracking. In our tracking stage, a hierarchical patch-based histogram is used to describe the observation model, computed by an improved L_1 bin-ratio dissimilarity (L_1 -BRD) distance. While in the detecting stage, a patch-based binary feature is obtained through center-symmetric local binary pattern (CS-LBP) and then used to train a randomize fern forest. We combine the two parts collaboratively and experiments demonstrate that the proposed tracking framework outperforms the state-of-the-art methods in challenging scenarios.

Keywords: hierarchical patch-based histogram, particle filter, patch-based binary feature.

1 Introduction

Object tracking is an important branch of computer vision, which has many practical applications in a range of fields, such as surveillance, intelligent traffic system, and medical imaging. Recently, more attention is focused on real-world tracking problems rather than lab-presumed environmental ones. These problems have increased difficulties in tracking since real-world sequences often suffer more from deformation, occlusion, and illumination changes.

Among the traditional tracking methods, particle filter [1,2] has represented good performance on handling non-gaussianity and multi-modality because we can design an accurate and fast recursive Bayesian model by using the particle filter. Brasnett *et al.* [3] integrated color, edge and texture cues into a particle filter with adaptive parameters. In [4], target candidate is sparsely represented in the space spanned by both target templates and trivial templates. A local patch-based appearance model is also introduced to describe the topology of the object by on-line update [5]. The robustness of each patch is determined by analyzing the likelihood landscape of the entire patch set. Meanwhile, sampling is an important approach to solve particle degeneration, recent study mainly focused on developing Monte Carlo sampling [6].

On the other hand, applying online learning to train a discriminative object detector during tracking has become a prevalent trend. In this setting a detector is trained using the samples from scratch in the first frame for any arbitrary objects that is desired to be tracked, and then the current tracker state is used to update the classifier. This leads to no restriction on the type of tracking object because no priori knowledge is attained before tracking. Bastian *et al.* [7] presented an approach for multi-object tracking which considers object detection and spacetime trajectory estimation as a coupled optimization problem. Babenko *et al.* [8] adopt an Online Multiple Instance Learning algorithm in their tracking system, which updates the appearance model with a set of image patches, switching the problem of sample selection from the tracking application towards the learning algorithm. Grabner *et al.* [9] formulated the updating process in a semi-supervised fashion since unlabeled data is explored in a principled manner. This has allowed to avoid the drifting problem when each updating of the tracker may introduce an error and accumulate. Tracking, learning and detection is investigated by Kalal *et al.* [10,11] in long-term tracking of unknown objects. They use a PN-Learning constrain to establish an incremental classifier and an adaptive median-flow tracker.

The idea of multi-scale feature extraction has been employed in detection and recognition [12]. Motivated by this, we construct an observation model relying on hierarchically-divided HSV histograms in the tracking section, which are composed of subregions with different sizes. Then we propose a new likelihood computation method based on L_1 -BRD, attaching different weights to the distance of each subregion. In this paper, we try to search a patch-based feature to train a randomized forest [13] for detection. Motivated by CS-LBP [14], a good descriptor based on comparison, we propose a new patch-based binary feature that has several advantages such as robustness to illumination changes and occlusion. L_1 -BRD[15] has been employed to describe the dissimilarity between histograms. At last, tracking and detection are implemented collaboratively to achieve satisfying performance on deformable object tracking, as will be illustrated in the following.

2 Related Work

2.1 Particle Filtering

Particle filtering employs a sequential Monte Carlo technique to solve Bayesian filtering, being aimed at approximating to approximate the posterior probability density of a state using a series of random particles with associated weights. It can be described as $X = \{(x_1, w_1), (x_2, w_2), \dots, (x_N, w_N)\}$, N is the total number of particles. In Sequential Importance Sampling (SIS) [16], the posterior probability density $p(x_k | z_{1:k})$ is:

$$p(x_k | z_{1:k}) \approx \sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)}) \quad (1)$$

where $w_k^{(i)}$ is the normalized weight of the i -th particle. In iteration, $w_k^{(i)}$ is updated by

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(z_k | x_k^{(i)}) p(x_k^{(i)} | x_{k-1}^{(i)})}{q(x_k^{(i)} | x_{k-1}^{(i)}, z_{1:k})} \quad (2)$$

$p(z_k | x_k^{(i)})$ is the observation likelihood density, measuring how much the target object and observation at the proposed state coincide. $p(x_k^{(i)} | x_{k-1}^{(i)})$ is the state transition density representing the next state x_k based on the previous state x_{k-1} . $q(x_k^{(i)} | x_{k-1}^{(i)}, z_{1:k})$, the importance sampling function, is always set to be ideally close to the posterior density. Hence, (2) could be rewritten as:

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(z_k | x_k^{(i)}) \quad (3)$$

When the tracking goes on, particles may become degenerate after several iterations, that is why we should introduce resampling. Resampling is an important step to eliminate low weight particles and regenerate a new set of particle with higher weight to guarantee the accuracy of the prediction and tracking results. In conclusion, particle filter operates based on four steps: target state generation, particle propagation, particle weight computation and final resampling.

2.2 Boosting Binary Classifier

Due to its speed, accuracy and possibility of incremental update, the randomized forest classifier [8] has attracted much attention in recent years. As mentioned in [17], the classifier is composed of a number of ferns [18] that are evaluated parallelly on each single patch. Figure 1 illustrates how the classifier works:



Fig. 1. A detector based on randomized fern forest classifier. The feature selected is in binary pattern.

Firstly, a Binary Pattern is introduced to represent the feature used in classifier. The used binary criterion are defined as

$$f_i = \begin{cases} 1 & \text{if } I(a_{i,1}) < I(a_{i,2}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $I(a_{i,1})$ and $I(a_{i,2})$ denote the sum of pixel intensity of the two parts respectively where the patch is divided horizontally and vertically. Each image patch is expressed with a number of local binary value where a series of distinguished position, size, ratio are generated randomly for each fern, resulting in a discrete vector when a patch comes in.

Then the leaf node of each fern stands for posterior probability according to the feature vector of the patch, being estimated incrementally throughout the whole learning process. Each leaf node records the number of positive p and negative n examples that fall into in during training and updating. The posterior of vector x can be estimated by maximize likelihood

$$\Pr(y=1|x) = \frac{p}{p+n} \quad (5)$$

The posteriors from all ferns are averaged and the classifier marks the image with a positive response if the average is larger than 50%.

3 Improved Tracking Methodology

3.1 A Novel Observation Model

In particle filter tracking, the observation model measures how much the target template and observation at the proposed state relate to each other. The discrimination of observation model has a big influence on tracking result. BRD [15], which is robust to normalization and partial matching problems, is used for the basis of our similarity measurements of the observation model.

An object is represented by a patch-based model as shown in Figure 2. The object is divided on 1×1 , 2×2 , 4×4 subregions respectively, so the observation model can

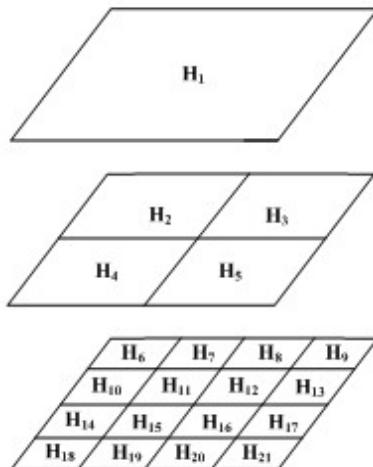


Fig. 2. Observation mode $\mathbf{Z} = (\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_{21})$, where \mathbf{H}_1 is the histogram of the whole image, \mathbf{H}_2 to \mathbf{H}_5 is histogram of 2×2 subregions, \mathbf{H}_6 to \mathbf{H}_{21} is histogram of 4×4 subregions

be described as $(\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_{21})$ where \mathbf{H}_i is the HSV-based histogram of the i -th patch. Assume the template target is $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{21})$ and the candidate target is depicted as $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_{21})$, to measure the similarity of \mathbf{p}_i and \mathbf{q}_i , we employ L_1 -BRD, which is a combination of BRD(Bin-Ratio Dissimilarity) with the widely used L_1 distance.

$$d_{l_{br}}(\mathbf{p}_i, \mathbf{q}_i) = \|\mathbf{p} - \mathbf{q}\|_1 - \|\mathbf{p} + \mathbf{q}\|_2^2 \frac{|p_m - q_m| p_m q_m}{(p_m + q_m)^2} \quad (6)$$

where p_m and q_m is the m -th bin in the i -th histogram. Considering the cases of deformation or occlusion in part of the candidate, a Gaussian kernels[19] weight is multiplied to each $d_{l_{br}}(\mathbf{p}_i, \mathbf{q}_i)$ to distinguish the influence of each histogram. The entire distance between template target and candidate target can be as follows:

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^N w_i d_{l_{br}}(\mathbf{p}_i, \mathbf{q}_i) \quad (7)$$

$$w_i = \frac{\exp(-\frac{1}{T} d_{l_{br}}(\mathbf{p}_i, \mathbf{q}_i))}{\sum_{j=1}^N \exp(-\frac{1}{T} d_{l_{br}}(\mathbf{p}_j, \mathbf{q}_j))}$$

so the observation likelihood density can be written as:

$$p(z_k | x_k^i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1-d(\mathbf{p}, \mathbf{q})}{2\sigma^2}\right) \quad (8)$$

where σ is a user-defined parameter. Because the observation model is constructed by a hierarchically divided subregions and the likelihood computation of each subregion is combined with a weight based on the distance between corresponding histogram, the model can describe the object more robustly, including being tolerate to deformation, occlusion in some degree.

3.2 Binary RCS-LBP

While the integral image used for binary pattern has an advantage over standard LBP by encoding the patch with fewer code, it has lost information on some important description of the image, such as texture. It should be noted that we can also convert the LBP descriptor into a binary form.

Unlike the traditional LBP, CS-LBP [14] adopts only center-symmetric pairs of pixels when computing LBP operator. The computation of CS-LBP is:

$$CS-LBP_{R,N,T}(x, y) = \sum_{i=0}^{(N/2)-1} s(n_i - n_{i+(N/2)}) 2^i, \quad (9)$$

$$S(x) = \begin{cases} 1 & x > T \\ 0 & otherwise \end{cases}$$

where n_i and $n_{i+(N/2)}$ correspond to the gray-values of center-symmetric pairs of pixels of N equally spaced pixels on a circle of radius R . And we define the rotation-invariant CS-LBP (RCS-LBP) by

$$RCS - LBP = \min \{CS - LBP_{R,N,T}, k \mid k = 0, 1, \dots, (N/2) - 1\} \quad (10)$$

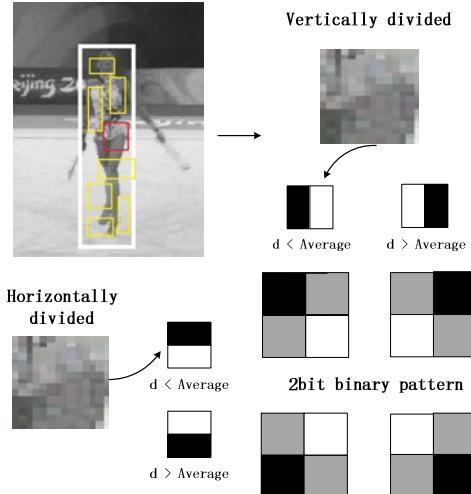


Fig. 3. Binary feature based on RCS-LBP

No-overlapping patches in the bounding box are generated with a measurement of Hessian Matrix. Horizontally or vertically partitioning each patch into two parts, we compute the L_2 distance of RCS-LBP based histogram between the two parts (Figure 3). Then the binary feature is obtained by

$$f_i = \begin{cases} 1 & \text{if } d_{l_2}(a_{i,1}, a_{i,2}) < \frac{1}{N} \sum_{i=1}^N d_{l_2}(a_{i,1}, a_{i,2}) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $d_{l_2}(a_{i,1}, a_{i,2})$ returns the L_2 distance of the two parts and N is the total number of patches. Then the feature vector is delivered to the randomized fern forest for training or detecting.

4 Fusion Algorithm

To overcome the drawbacks of traditional tracking methods and improve the robustness, in this paper, we propose a novel joint tracking and detection method, which has a judgment on the result of pure tracking and can detect the object if tracking fails. Besides, it updates the template model adaptively according to the score of the classifier. The flow chart is demonstrated in Figure 4.

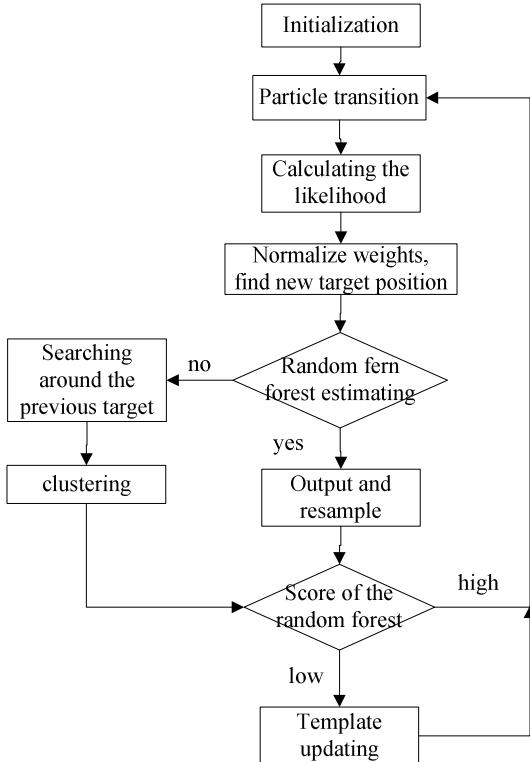


Fig. 4. The framework of the proposed method

Firstly, an estimated target is traced relying on the particle filter tracking algorithm using the proposed observation likelihood density. Then we use the trained randomized fern forest to discriminate the label of the estimated target. If the target is labeled positive, the estimated is considered to be the real position of the target and the averaged posterior of all ferns is recorded.

On the contrary, when the traced target is labeled negative by the classifier, we can infer that there's a failure in the tracking so a scanning strategy is promoted in a relative circus around the target in the previous frame. After clustering, a new target position is located and it is also delivered to the randomized forest to get a score.

Finally, in case of changes in object appearance, the score of the randomized forest is used to update the target template M_t ,

$$M_{t+1} = \begin{cases} M_{t+1} = \beta M_t + (1-\beta) M_{int} & \beta < T \\ M_{t+1} = M_t & \text{otherwise} \end{cases} \quad (12)$$

where β is the averaged posterior $\Pr_{avg}(y=1|x)$. M_{int} indicates the local patch model in the first frame which prevents from losing enormous information on the original and M_t represents the model image obtained in the previous frame. T is a pre-defined threshold which controls our update strategy. When the averaged posterior is higher than this threshold, the target is considered to be slightly changing its appearance and

the template needs update. Otherwise, the target is more likely to be occluded and the template should be updated adaptively, so are the classifier and patch-based histogram.

5 Experimental Results

We test our tracking method on some challenging sequences, and some of them are publicly available. Objects in the two tracking sequences in Figure 5 suffer from deformation on appearance. In Figure 5(a)-(e), the face of the student suffers from different degree and orientation of occlusion. From the tracking sequences, we can see that both the particle filter and TLD cannot track the object precisely, existing some deviation on certain frames. Thanks to the patch-based method we use in our integrated tracking methodology, our tracker is more robust against the occlusion and movement of the target and thus produce a more reliable tracking result.



Fig. 5. Tracking results of two sequences. The first row of each sequence is obtained by traditional particle filter, the second row is obtained by TLD, and the results of the third row are got from our proposed method.

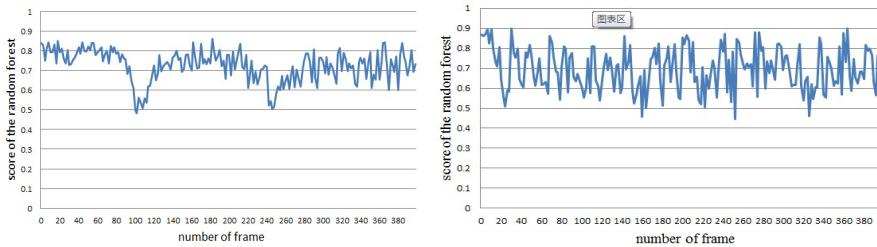


Fig. 6. The score of randomized fern forest of the two tracking sequences

In Figure 5(f)-(j), the dancer has relatively dramatic changes in body postures, accompanied with some movements in position. As shown in Figure 6(a), the score of randomized forest fluctuate all the time due to the continued change in appearance, and for several times the score is below 50% so detection is adopted. Particle filter tracking can roughly follow the trajectory of the dancer, but it deviate the accurate position sometimes. While TLD tracking loses the target when there are unexpected changes in the moving direction, as shown in frame 246, although it can re-detect the object through global searching. Compared with the two methods, our tracker is not influenced by the deformation and can give the exact position of the object.

6 Conclusion

In this paper, we propose and demonstrate an effective and robust tracking method based on the collaboration of tracking and detecting. Both the tracking and the detecting section are using a patch-based ideology. A novel binary RCS-LBP feature is adopted to train and update the randomized fern forest that is more tolerate to changes of the object. With respect to the observation model in particle filter tracking, we divide the region hierarchically to get a multidimensional HSV histogram, then the likelihood is computed by a weighted L_1 -BRD method which enables our tracker to better handle deformation and occlusion. The contributions of these new features detecting and hierarchical histogram particle filter are integrated in a unified tracking manner. The experimental results demonstrated that the proposed method performs robustly in various scenarios.

Acknowledgement. This paper is sponsored by the National Science Foundation of China under Grant No.61075012.

References

1. Ristic, B., Arulampalm, S., Gordon, N.: Beyond the Kalman filter: Particle filters for tracking applications. Artech House Publishers (2004)
2. Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: fifteen years later. Handbook of Nonlinear Filtering 12, 656–704 (2009)

3. Brasnett, P., Mihaylova, L., Bull, D., et al.: Sequential Monte Carlo tracking by fusing multiple cues in video sequences. *Image and Vision Computing* 25(8), 1217–1227 (2007)
4. Mei, X., Ling, H.: Robust visual tracking using ℓ_1 minimization. In: Proceedings of IEEE International Conference on Computer Vision (2009)
5. Kwon, J., Lee, K.: Highly Non-Rigid Object Tracking via Patch-based Dynamic Appearance Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013)
6. Kwon, J., Lee, K.M.: Tracking of abrupt motion using Wang-Landau Monte Carlo estimation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I*. LNCS, vol. 5302, pp. 387–400. Springer, Heidelberg (2008)
7. Leibe, B., Schindler, K., Cornelis, N., et al.: Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(10), 1683–1698 (2008)
8. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
9. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I*. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
10. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(7), 1409–1422 (2012)
11. Kalal, Z., Matas, J., Mikolajczyk, K.: Online learning of robust object detectors during unstable tracking. In: *IEEE 12th International Conference on Computer Vision Workshops* (2009)
12. Zhao, C., Liu, C., Lai, Z.: Multi-scale gist feature manifold for building recognition. *Neurocomputing* 74(17), 2929–2940 (2011)
13. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
14. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with center-symmetric local binary patterns. In: Kalra, P.K., Peleg, S. (eds.) *ICVGIP 2006*. LNCS, vol. 4338, pp. 58–69. Springer, Heidelberg (2006)
15. Xie, N., Ling, H., Hu, W., et al.: Use bin-ratio information for category and scene classification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2010)
16. MacEachern, S.N., Clyde, M., Liu, J.S.: Sequential importance sampling for nonparametric Bayes models: The next generation. *Canadian Journal of Statistics* 27(2), 251–267 (1999)
17. Kalal, Z., Matas, J., Mikolajczyk, K.: Pn learning: Bootstrapping binary classifiers by structural constraints. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
18. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
19. Chapelle, O., Haffner, P., Vapnik, V.N.: Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks* 10(5), 1055–1064 (1999)

Adaptive Regularization Parameters and Norm Selection for Sparse Gradient Based Image Restoration

Xinqian Lin, Hongzhi Zhang, Hong Deng, and Wangmeng Zuo

Biocomputing Research Centre, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin 150001, China

suminglinxinqian@sina.com, {zhanghz0451, cswmzuo}@gmail.com,
denghong_hit@163.com

Abstract. The performance of image restoration depends on the choice of image priors and regularization parameters. Recent studies reveal that heavy-tailed sparse gradient distributions of natural images can be modeled by hyper-Laplacian with the norm $p \in [0.5, 1]$. However, it is still very challenging to determine the values of the norm and the regularization parameters. In this paper, we proposed a maximum a posterior (MAP) method for adaptive regularization parameters and norm selection. During the procedure of image restoration, by modeling the estimated residual and gradient distribution with Gaussian and hyper-Laplacian, respectively, we suggest a MAP formulation for the joint estimation of the latent image, regularization parameter, and norm. Then, we propose an alternating optimization method to iteratively solving the MAP problem. Experimental results show that, the proposed method obtained satisfactory restoration results for various degraded images with different noise level, and outperformed the other methods.

Keywords: Regularization, image restoration, maximum a posterior (MAP), hyper-Laplacian.

1 Introduction

Image restoration aims to restore the latent clear $m \times n$ image \mathbf{x} from degraded or noisy observation \mathbf{y} , and is an inherent ill-posed problem which has been extensively studied. To solve this problem, one typical image restoration model usually includes two terms: fidelity term and regularizer. Based on natural image statistics, a number of methods have been developed for modeling image priors and regularizers. One representative class of image priors is the gradient priors based on the observation that natural images generally have a heavy-tailed distribution of gradients. The use of gradient prior can be traced back to 1990s, when Rudin et al. [1] proposed a total variation (TV) model for image denoising, where the gradients are actually modeled as Laplacian distribution. Besides, mixture of Gaussians (GMM) can also be used to approximate the distribution of gradient magnitude [2, 3]. Recently, hyper-Laplacian model was proposed to model the heavy-tailed distribution of gradients, and has been widely applied to various image restoration tasks [4, 5, 6, 7].

Using the hyper-Laplacian prior, the image restoration model can be formulated as,

$$\mathbf{x} = \arg \min_{\mathbf{x}} \frac{\|\mathbf{Ax} - \mathbf{y}\|_2^2}{2\sigma^2} + \alpha \|\mathbf{Dx}\|_p^p \quad (1)$$

where \mathbf{A} is the degradation operator, $\lambda = \alpha\sigma^2$ is the regularization parameter, p is the norm, and $\mathbf{D} = \{\mathbf{D}_h, \mathbf{D}_v\}$ denotes the gradient operator in the horizontal and vertical directions, respectively. Given p and λ , several algorithms had been proposed to solve the sparse gradient based image restoration problem in Eq. (1) [4, 5].

However, different clear images have different p and α values. As shown in Fig. 1, the p value of images with rich textures (Fig. 1(a)) is higher than that of smooth images (Fig. 1(c)). Moreover, the noise level of the degraded image would also affect the value of regularization parameter. Actually, the restoration performance depends on the value of the norm and the regularization parameters, but by far little attentions was given in this topic. Thus, if the fixed norm and regularization parameter are adopted for the degraded images, the model in Eq. (1) would fail in some cases.

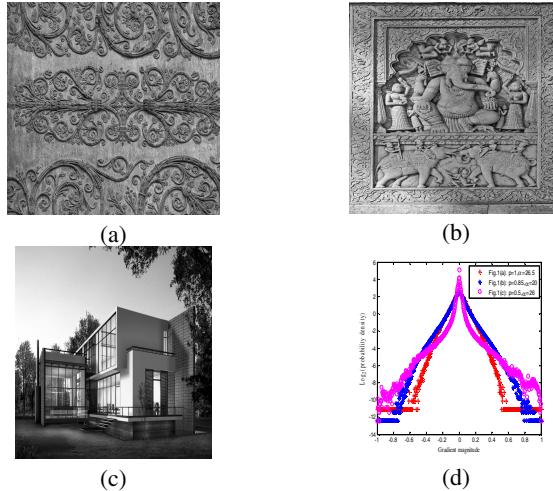


Fig. 1. Latent images and their distributions of gradient magnitudes

By far, several parameter selection methods had been proposed. As to noise level estimation, several methods were proposed to estimate the standard deviation of additive white Gaussian noise (AWGN) [8, 9]. As to regularization parameter selection, the existing methods, e.g., generalized cross-validation (GCV) [10], L-curve [11], and discrepancy principle [12], were mainly designed for TV-based image restoration, but little attentions were given on the adaptive regularization parameter and norm selection for hyper-Laplacian based image restoration.

In this paper, using the MAP framework, we present a probabilistic framework which allows the joint estimation of the latent image, regularization parameter, and norm. During the procedure of image restoration, we propose an alternated optimization method for adaptive regularization parameters and norm selection. Thus,

we can dynamically tune the regularization parameters and norm, and can obtain satisfactory restoration results for degraded images under different noise levels and sparse gradient priors. Experimental results show that, the proposed method is better than the method with the fixed hyper-Laplacian prior.

2 Probabilistic Image Restoration with Hyper-laplacian Prior

In this section, we first discuss the degradation model and the hyper-Laplacian based image prior model. Then, we formulate image restoration in the maximum a posterior (MAP) framework. Different from [5], our formulation allows the joint estimation of the latent image \mathbf{x} , the regularization parameter, and the norm p .

For non-blind image restoration, the blur kernel \mathbf{k} is assumed to be known in advance. The degraded observation \mathbf{y} can then be modeled by,

$$\mathbf{y} = \mathbf{k} \otimes \mathbf{x} + \mathbf{n}, \quad (2)$$

where \mathbf{n} is the additive white Gaussian noise (AWGN) with the standard deviation of σ , and \otimes denotes the convolution operator. First, the gradients of the latent image \mathbf{x} is modeled with the hyper-Laplacian distribution,

$$q(\mathbf{x} | \alpha, p) = \prod_i \frac{p}{2} \left(\frac{\alpha}{2} \right)^{\frac{1}{p}} \frac{1}{\Gamma(\frac{1}{p})} \exp\left(-\frac{\alpha}{2} |\mathbf{D}\mathbf{x}|_p^p\right), \quad (3)$$

where $\Gamma(\cdot)$ is the Gamma function, α and p are the parameters to control the shape of the hyper-Laplacian distribution, and $\mathbf{D} = \{\mathbf{D}_h, \mathbf{D}_v\}$ denotes the gradient operator with $\mathbf{D}_h\mathbf{x}$ and $\mathbf{D}_v\mathbf{x}$ stand for the gradients of the image in the horizontal and vertical directions, respectively. Second, because \mathbf{n} is AWGN, it is natural to model $q(\mathbf{y} | \mathbf{x}, \mathbf{k}, \sigma^2)$ with the Gaussian distribution,

$$q(\mathbf{y} | \mathbf{x}, \mathbf{k}, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{k} \otimes \mathbf{x})_i^2}{2\sigma^2}\right), \quad (4)$$

where σ denotes the standard deviation of the Gaussian distribution.

Based on Bayes theorem, the posterior probability $q(\mathbf{x}, \sigma^2, \alpha, p | \mathbf{y}, \mathbf{k})$ is proportional to the product of $q(\mathbf{y} | \mathbf{x}, \mathbf{k}, \sigma^2)$ and $q(\mathbf{x} | \alpha, p)$,

$$q(\mathbf{x}, \sigma^2, \alpha, p | \mathbf{y}, \mathbf{k}) \propto q(\mathbf{y} | \mathbf{x}, \mathbf{k}) \cdot q(\mathbf{x} | \alpha, p). \quad (5)$$

Thus the MAP estimation $MAP_{x, \sigma^2, \alpha, p}$ of \mathbf{x} , σ^2 , α and p can be formulated as,

$$\begin{aligned} \mathbf{x}, \sigma^2, \alpha, p &= \arg \min_{\mathbf{x}, \sigma^2, \alpha, p} \left\{ -\log q(\mathbf{x}, \sigma^2, \alpha, p | \mathbf{y}, \mathbf{k}) \right\} \\ &= \arg \min_{\mathbf{x}, \sigma^2, \alpha, p} \left\{ \begin{aligned} &\frac{1}{2} mn \log(2\pi\sigma^2) + \frac{\|\mathbf{y} - \mathbf{k} \otimes \mathbf{x}\|^2}{2\sigma^2} + \frac{\alpha}{2} \|\mathbf{D}\mathbf{x}\|_p^p \\ &+ (2mn - m - n) \left(\log(\Gamma(\frac{1}{p})) - \log(\frac{p}{2}) - \frac{1}{p} \log(\frac{\alpha}{2}) \right) \end{aligned} \right\}. \end{aligned} \quad (6)$$

3 Adaptive Regularization Parameter and Norm Selection

In this section, we propose an alternating optimization method for the estimation of the latent image \mathbf{x} , the standard deviation σ , and the hyper-Laplacian parameters α and p . Generally, the proposed method iteratively solves two subproblems: 1) the updating of \mathbf{x} , and 2) the updating of σ , α , and p . In the following, we introduce the solutions to these two subproblems in detail.

3.1 The Updating of \mathbf{x}

Given σ , α , and p , we define $\lambda = \alpha\sigma^2/2$, and the problem in Eq. (6) becomes,

$$\mathbf{x} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{k} \otimes \mathbf{x}\|_2^2 + \lambda \|\mathbf{Dx}\|_p^p. \quad (7)$$

Here we adopt the fast deconvolution algorithm in [5] to solve this subproblem. By introducing two auxiliary variables \mathbf{w}_h and \mathbf{w}_v , we use the half-quadratic penalty method [5, 13, 14], and introduce a relaxed formulation of the problem in Eq. (7):

$$\arg \min_{\mathbf{x}, \mathbf{w}_h, \mathbf{w}_v} \frac{1}{2} \|\mathbf{y} - \mathbf{k} \otimes \mathbf{x}\|_2^2 + \frac{\gamma}{2} \left(\|\mathbf{D}_h \mathbf{x} - \mathbf{w}_h\|^2 + \|\mathbf{D}_v \mathbf{x} - \mathbf{w}_v\|^2 \right) + \lambda (\|\mathbf{w}_h\|_p^p + \|\mathbf{w}_v\|_p^p). \quad (8)$$

When the penalty coefficient $\gamma \rightarrow \infty$, the solution of Eq. (8) would converge to that of Eq. (7). In practice, we adopt the continuation technique by initializing γ with a smaller value and gradually increasing it until convergence.

Then, we alternatively update \mathbf{x} , \mathbf{w}_h and \mathbf{w}_v by solving the following subproblems:

(1) Updating \mathbf{x} : Given \mathbf{w}_h and \mathbf{w}_v , we update \mathbf{x} by solving the quadratic problem

$$\arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{k} \otimes \mathbf{x}\|_2^2 + \frac{\gamma}{2} \left(\|\mathbf{D}_h \mathbf{x} - \mathbf{w}_h\|^2 + \|\mathbf{D}_v \mathbf{x} - \mathbf{w}_v\|^2 \right). \quad (9)$$

The closed-form solution of \mathbf{x} can be written as,

$$\mathbf{x} = \left(\frac{1}{\gamma} \mathbf{K}^T \mathbf{K} + \mathbf{D}_h^T \mathbf{D}_h + \mathbf{D}_v^T \mathbf{D}_v \right)^{-1} \left(\frac{1}{\gamma} \mathbf{K}^T \mathbf{y} + \mathbf{D}_h^T \mathbf{w}_h + \mathbf{D}_v^T \mathbf{w}_v \right) \quad (10)$$

where \mathbf{K} is the matrix representation of the blur kernel \mathbf{k} . With the fast Fourier transform (FFT), the solution of \mathbf{x} can be computed more efficiently by,

$$\mathbf{x} = FFT^{-1} \left(\frac{\overline{FFT(\mathbf{k})} \circ FFT(\mathbf{y}) / \gamma + FFT(\mathbf{D}_h)^* \circ FFT(\mathbf{w}_h) + FFT(\mathbf{D}_v)^* \circ FFT(\mathbf{w}_v)}{\overline{FFT(\mathbf{k})} \circ FFT(\mathbf{k}) / \gamma + \overline{FFT(\mathbf{D}_h)} \circ FFT(\mathbf{D}_h) + \overline{FFT(\mathbf{D}_v)} \circ FFT(\mathbf{D}_v)} \right), \quad (11)$$

where “ \circ ” denotes entry-wise multiplication, and “ $-$ ” denotes the entry-wise division.

(2) Updating \mathbf{w}_h and \mathbf{w}_v : Given \mathbf{x} , the sub-problem of updating \mathbf{w}_h and \mathbf{w}_v becomes,

$$\arg \min_{\mathbf{w}_h, \mathbf{w}_v} \frac{1}{2} \left(\|\mathbf{w}_h - \mathbf{w}_h^0\|^2 + \|\mathbf{w}_v - \mathbf{w}_v^0\|^2 \right) + \frac{\lambda}{\gamma} (\|\mathbf{w}_h\|_p^p + \|\mathbf{w}_v\|_p^p), \quad (12)$$

where $\mathbf{w}_h^0 = \mathbf{D}_h \mathbf{x}$ and $\mathbf{w}_v^0 = \mathbf{D}_v \mathbf{x}$. The problem in Eq. (12) can be decomposed into $(2mn - m - n)$ independent 1D optimization subproblems,

$$\arg \min_w \frac{1}{2} \|w - w_0\|^2 + \frac{\lambda}{\gamma} |w|_p^p. \quad (13)$$

For each 1D subproblem, we use the lookup table (LUT) method described in [5] to solve it.

3.2 The Updating of σ , α , and p

Given \mathbf{x} , let $r^2 = \|\mathbf{y} - \mathbf{k} \otimes \mathbf{x}\|^2$ and $\mathbf{d} = \mathbf{Dx}$, and the optimization problem in Eq. (6) can be reformulated as

$$\arg \min_{\sigma^2, \alpha, p} \left\{ \begin{aligned} & \frac{1}{2} mn \log(2\pi\sigma^2) + \frac{r^2}{2\sigma^2} \\ & + \frac{\alpha}{2} \|\mathbf{d}\|_p^p + (2mn - m - n) \left(\log(\Gamma(\frac{1}{p})) - \log(\frac{p}{2}) - \frac{1}{p} \log(\frac{\alpha}{2}) \right) \end{aligned} \right\}, \quad (14)$$

which can be further decomposed into two simple subproblems:

$$\arg \min_{\sigma^2} \left\{ \frac{1}{2} mn \log(\sigma^2) + \frac{r^2}{2\sigma^2} \right\}, \quad (15)$$

$$\arg \min_{\alpha, p} \left\{ \frac{\alpha}{2} \|\mathbf{d}\|_p^p + (2mn - m - n) \left(\log(\Gamma(\frac{1}{p})) - \log(\frac{p}{2}) - \frac{1}{p} \log(\frac{\alpha}{2}) \right) \right\}. \quad (16)$$

By solving the problem in Eq. (15), we obtain a closed-form solution of σ^2 ,

$$\sigma^2 = \frac{r^2}{mn}. \quad (17)$$

The solution to the problem in Eq. (16) is more difficult. We observe that, given p , the problem in Eq. (16) becomes,

$$\arg \min_{\alpha} \left\{ \frac{\alpha}{2} \|\mathbf{d}\|_p^p + (2mn - m - n) \frac{1}{p} \log(\frac{\alpha}{2}) \right\}, \quad (18)$$

and the closed-form solution on α is,

$$\alpha = \frac{2mn - m - n}{p \|\mathbf{d}\|_p^p}. \quad (19)$$

By substituting Eq. (19) in to Eq. (16), we use a simple 1D exhaustive searching strategy to obtain the optimal estimation on p .

4 Experimental Results

In this section, experiments are conducted to evaluate the restoration performance of the proposed MAP-based method. As shown in Fig. 2, we use six clear images and two blur kernels to test the proposed method. For each image, we test the performance of the restoration method under two noise levels: $\sigma = 0.01$ and $\sigma = 0.001$. Based on signal noise ration (SNR), we compare the proposed method with the fast deconvolution (FastDeconv) method with the fixed $p = 2/3$ [5], and the fast TV-based restoration (FTVd) method [14]. For each of FastDeconv and FTVd, we implement two variants: 1) Assuming that the noise level is known, we set the regularization parameter by $\lambda = 20\sigma^2$; 2) Using the method in [9] to estimate the noise level $\hat{\sigma}^2$, we set the regularization parameter by $\lambda = 20\hat{\sigma}^2$.

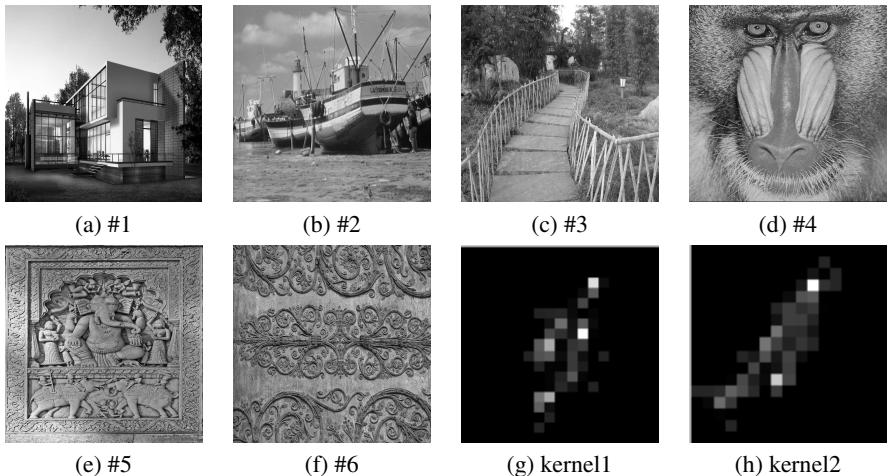
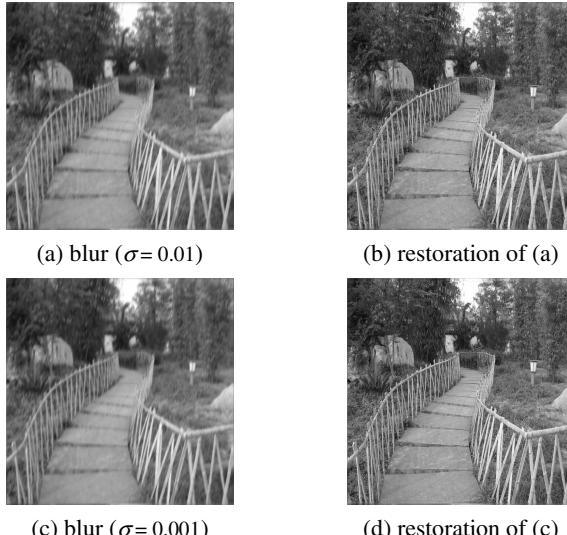


Fig. 2. Images and kernels used in our experiments. a-f) are the images; g-h) are the kernels

Table 1 lists the SNR values of FastDeconv, FTVd, and the proposed method by using kernel 1 to blur the clear images. The SNR values of the first FastDeconv variant are listed in the left column, while those of the second variant are listed in the right column. So do FTVd. One can see that, hyper-Laplacian prior generally is better than TV-based model for image restoration, and the proposed MAP method usually achieves higher SNR values. Note that in FastDeconv and FTVd, even we assume the correct noise level is known, the proposed method can achieve similar or higher SNR values. We can observe the similar results based on the SNR values of FastDeconv, FTVd, and the proposed method by using kernel 2. Thus, we conclude that, the proposed MAP method is effective for the joint estimation of \mathbf{x} , λ , and p .

Table 1. The SNR values of the restoration results with blur kernel 1

Image#	$\sigma = 0.01$			$\sigma = 0.001$						
	FastDeconv [5]	FTVd [14]	MAP	FastDeconv [5]	FTVd [14]	MAP				
1	18.23	17.64	17.38	15.51	18.06	26.23	22.33	24.74	24.18	27.27
2	15.79	15.42	14.30	12.97	15.27	22.57	23.66	21.14	22.34	23.61
3	12.71	12.85	12.55	12.12	12.72	21.03	21.80	20.07	21.33	21.77
4	9.78	10.07	10.08	9.80	10.06	18.35	18.70	17.58	18.07	18.30
5	10.76	10.91	11.08	10.14	10.99	19.21	14.69	18.50	16.77	19.81
6	8.94	9.22	9.45	9.28	9.41	17.89	17.80	17.08	18.66	18.69
Average	12.70	12.69	12.48	11.64	12.75	20.88	19.83	19.85	20.23	21.58

**Fig. 3.** Blur images and restoration results obtained using the proposed method

Finally, using image #3 and kernel 2, Fig. 3 shows the blur images and the restoration results of the proposed method. From Fig. 3, one can see that the proposed method is effective for the restoration of images with different noise levels.

5 Conclusion

Image prior and regularization parameter play a critical role in image restoration. Adaptive choice of regularization and image prior parameters can improve both the flexibility and restoration performance of the restoration model. Based on hyper-Laplacian prior, we proposed a MAP formulation of the image restoration model which allows the joint estimation of the latent image, regularization parameter, and

norm. Then, we developed an efficient alternated optimization scheme for solve the MAP model. Finally, experimental results showed that the proposed method is effective for adaptive regularization parameter and norm selection. In the future, we will extend the proposed method to estimate these parameters locally, and investigate more automated approaches for both blind and non-blind image restoration.

Acknowledgment. This work is supported by NSFC under Grant No.s 61271093 and 61001037, and the program of ministry of education for new century excellent talents.

References

1. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear Total Variation Based Noise Removal Algorithms. *J. Physica D* 60(1-4), 259–268 (1992)
2. Fergus, R., Singh, B., Hertzmann, A., Roweis, S., Freeman, W.T.: Removing Camera Shake from a Single Photograph. *Proc. ACM* 25(3), 787–794 (2006)
3. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Efficient Marginal Likelihood Optimization in Blind Deconvolution. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2657–2664. IEEE (2011)
4. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and Depth from a Conventional Camera with a Coded Aperture. *ACM* 26(3), 70 (2007)
5. Krishnan, D., Fergus, R.: Fast Image Deconvolution Using Hyper-Laplacian Priors. In: *NIPS*, vol. 22, pp. 1–9 (2009)
6. Cho, T.S., Joshi, N., Zitnick, C.L., et al.: A Content-Aware Image Prior. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 169–176 (2010)
7. Cho, T.S., Zitnick, C.L., Joshi, N., et al.: Image Restoration by Matching Gradient Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4), 683–694 (2012)
8. Donoho, D.L.: De-Noising by Soft-Thresholding. *IEEE Transactions on Information Theory* 41(3), 613–627 (1995)
9. Stefano, A., White, P.R., Collis, W.B.: Training Methods for Image Noise Level Estimation on Wavelet Components. *J. EURASIP on Advances in Signal Processing* 16, 2400–2407 (2004)
10. Galatsanos, N.P., Katsaggelos, A.K.: Methods for Choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and their Relation. *IEEE Transactions on Image Processing* 1(3), 322–336 (1992)
11. Hansen, P.C.: Analysis of Discrete Ill-posed Problems by Means of the L-curve. *SIAM Review* 34(4), 561–580 (1992)
12. Wen, Y.W., Chan, R.H.: Parameter Selection for Total-Variation-Based Image Restoration Using Discrepancy Principle. *IEEE Transactions on Image Processing* 21(4), 1770–1781 (2012)
13. Geman, D., Yang, C.: Nonlinear Image Recovery with Half-quadratic Regularization. *IEEE Transactions on Image Processing* 4(7), 932–946 (1995)
14. Wang, Y., Yang, et al.: A New Alternating Minimization Algorithm for Total Variation Image Reconstruction. *SIAM Journal on Imaging Sciences* 1(3), 248–272 (2008)

Key-Frame Selection Strategy Based on Edge Points Classification in 2D-to-3D Conversion

Jiangchuan Xie¹, Jiande Sun^{1,2,*}, Ju Liu^{1,2}, and Qiaoli Hu¹

¹ School of Information Science and Engineering, Shandong University, Jinan 250100, China

² The Hisense State Key Laboratory of Digital-Media Technology, Qingdao 266061, China
{Jiejiangchuan, hugiaolichola}@126.com, {jd_sun, juliu}@sdu.edu.cn

Abstract. In 2D-to-3D video conversion, key-frame selection is essential and it affects the quality and workload of the conversion. In this paper, a key-frame selection method based on edge classification is proposed. In the proposed method, the candidate key-frames are selected out based on the occlusion area and feature point correspondence. The optimal key-frames are selected out from the candidates according to the edge point classification so that the depth of the key-frames can be estimated accurately and automatically referring to the depth estimation based on focus cue. Experiments show that the proposed method can bring 2D-to-3D conversion better objective and subject quality and it is promising in practical applications.

Keywords: Key-frame Selection, Edge Points Classification, 2D-to-3D Conversion, Focus Cue, Depth Estimation.

1 Introduction

Recently, 2D-to-3D conversion has attracted more and more attentions as it can produce 3D immersive feeling directly from 2D video.

Usually 2D-to-3D conversion contains the following steps: 1) key-frame selection, 2) depth estimation of key-frames, 3) depth propagation to non-key-frames and 4) virtual view synthesis, and generates the left and right view sequence by Depth Imaged Based Rendering (DIBR) [1]. At present, 2D-to-3D conversion can be divided into full-automatic and semi-automatic methods [2-5]. Full-automatic methods derive 3D video from 2D video directly without any manual interaction. Though the full-automatic method needs no human forces, it is not an optimal conversion as the quality of the generated 3D video is not high usually. Comparing with full-automatic 2D-to-3D conversion, the semi-automatic method selects key-frames and assigns depth information for key-frames manually. It can produce much higher 3D quality, so it has been more and more popular. However the selection of key-frames and the manual assignment of depth information for them need huge of labor cost. No matter for the full- or semi-automatic 2D-to-3D conversion methods, key-frame selection is the most critical issue that determines the quality of the generated 3D video as it is the

* Corresponding author.

first step in 2D-to-3D conversion. However, there are only few studies on how to select suitable key-frames in 2D-to-3D conversion.

Key-frame selection is a classical study in the field of video content representation and analysis [6-9]. Such methods are designed for video analysis, video retrieval and so on, but few are for 2D-to-3D conversion. In the 2D-to-3D conversion of [5], key-frames were selected at a fixed temporal interval. Nistér selected key-frames based on the sharpness of frames, which was the mean square of the horizontal and vertical derivatives [10]. Gibson selected the key-frames taking account of the error calculated from the fundamental matrix and the 2D projective transformation matrix [11]. The performance of these methods is not satisfied. In [12], the occlusion area and correspondence ratio were utilized to select candidate key-frames out firstly and the re-projection error was calculated to determine the optimal key-frames. However re-projection error is not a direct factor related to depth information.

In this paper, we propose a novel key-frame selection method to select out the optimal key-frames from the point of depth estimation. As far as our knowledge, it is the first time to select key-frames for 2D-to-3D conversion from the point of depth estimation. The proposed method can be concluded as follows: 1) Correspondence ratio and occlusion area calculation for candidate key-frame selection as in [12]; 2) Color difference and foreground object size analysis and candidate key-frames refreshing; 3) Edge points classification and optimal key-frames determination. The flow chart of the proposed method is shown in Fig. 1. Given the proposed method, the semi-automatic 2D-to-3D conversion can be turned into full-automatic conversion with little quality degradation under a certain depth estimation algorithm. The objective and subjective experiment results verify that the feasibility and reliability of the proposed method.

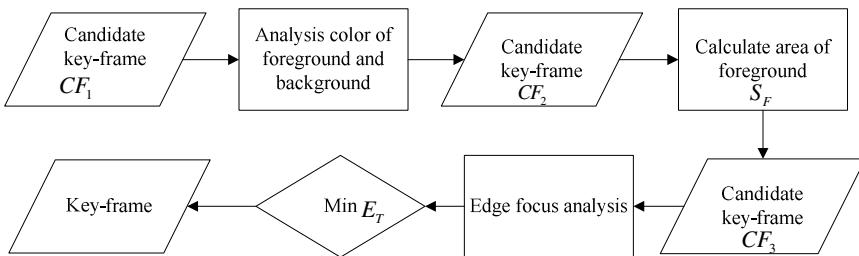


Fig. 1. The Flow Chart of the Proposed Method

2 Candidate Key-Frames Refreshing

When the preliminary candidate key-frames CF_1 are determined as in [12], the optimal ones are selected out based on depth estimation. In this paper, the depth estimation methods proposed in [13] and [14] are adopted, which estimate depth information based on wavelet transform. The number of nonzero coefficients in high frequency wavelet band, i.e. LH, HL and HH bands can reflect whether the details are focused or not and it can be used to produce the relative depth value for key-frames. Different from [13], in this paper, the localized spatial frequency is measured by

performing wavelet transform in the local window for each pixel and the size of each window is 16*16. Consequently the depth map generated here is not blocky as [14].

Given the method of depth estimation, the optimal key-frames can be selected out by refreshing the candidate key-frames based on color difference between background and foreground, foreground object size and edge point classification respectively.

2.1 Color Difference between Background and Foreground

If the foreground and background objects in a preliminary candidate key-frame have similar color or the foreground objects are textureless, the depth estimation of foreground object will be difficult. Such a candidate frame cannot be an optimal one.

Here the preliminary candidate frame CF_1 is segmented by using the Graph Cuts algorithm and pixels with similar color are assigned to the same cluster as in [15-17]. The frames, in which the colors of the foreground and background objects are similar, are kicked out from CF_1 . The candidate key-frames are refreshed to CF_2 .

2.2 Size of Foreground Object

Usually those outstanding foreground objects attract more attention than objects of background in a 3D video. That is to say, the stereoscopic feeling of viewers is mainly gained from the foreground objects. The viewers can easily figure out whether an object in a video is near or far from the camera if the background and foreground objects are well-spaced. Therefore, the area of foreground objects is a reasonable factor f selecting key-frames.

Usually the foreground objects are focused and with much more high frequency components, while the background objects are defocused and with much less high frequency components. In the focused foreground, the amount of nonzero coefficients is large and depth value estimated by wavelet transform is high. In the blurred background, the estimated depth value is low because a large number of high frequency coefficients are zero. Therefore, the frames with large foreground objects are better choice for key-frame selection in 2D-to-3D conversion.

The area of foreground objects of each frame in the candidate key-frames CF_2 is calculated. The frames with large foreground object area S_F are selected as new candidate key-frames CF_3 . The foreground object area S_F is the total number of pixels that belong to the foreground objects. S_A is the total pixel number in an image. New candidate key-frames CF_3 are selected as the following equations:

$$S_p^f = \frac{S_F^f}{S_A}, \quad f \in CF_2 \quad (1)$$

$$\text{if } S_p^f > S_Y, \text{ then } f \Rightarrow CF_3 \quad (2)$$

Where S_p^f is the area ratio of the frame f in the candidate key-frames CF_2 . S_Y is the area threshold, which belongs to $[0.2, 0.5]$.

The result of Graph Cuts segmentation is shown in Fig. 2(b). It can be seen from Fig. 2(a) that three kinds of colors exist in the image: 1) two fighting athletes, the audiences and the plant are black; 2) the mist is white; 3) the background is red.

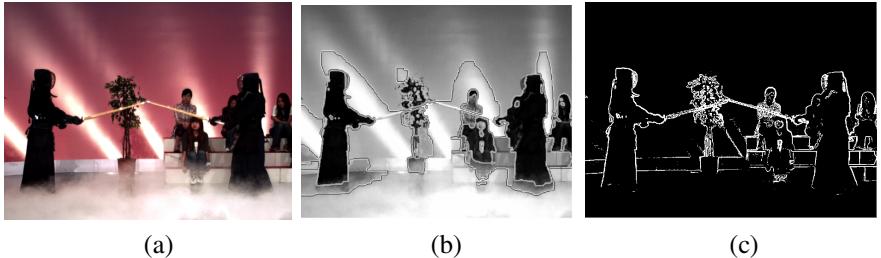


Fig. 2. (a) Color Image (b) Segmentation Result and (c) Edge Image

2.3 Edge Points Classification

An edge detector is implemented by using wavelet multi-resolution analysis [13]. Gaussian function is chosen as the smoothing function $\theta(x)$. The $\frac{d\theta(x)}{dx}$ is the first order of the function $\theta(x)$, and it can be seen as a wavelet because its integral in $(-\infty, +\infty)$ is equal to 0.

The definition of the dilation of $\theta(x)$ by a scaling factor is:

$$\theta_s(x) = \frac{1}{s} \theta\left(\frac{x}{s}\right) \quad (3)$$

The wavelet transform of $f(x)$ at the scale s is computed with the respect to the wavelet $\frac{d\theta_s(x)}{dx}$:

$$W_s f(x) = f(x) \otimes s \frac{d\theta_s(x)}{dx} \quad (4)$$

We detect edge points by identifying the maximum absolute value of the wave transforms by the following equation:

$$\left\{ \begin{array}{l} |W_s f(x_0)| > |W_s f(x_0+1)| \\ |W_s f(x_0)| > |W_s f(x_0-1)| \end{array} \right\} \quad (5)$$

An edge image is shown in Fig. 2(c). We classify the detected edge points into three categories. If a row of the edge image has n edge points, this row can be separated into $n+1$ row sections with different depth values. A high value of Lipschitz exponents ($0 < \alpha < 1$) of adjacent edge points means the variation between these edge points is small. The corresponding row section is considered defocused

and belonging to the background. We assign small depth value to this row section. Such edge points are regarded as the first kind edge points.

If the adjacent edge points have low value of Lipschitz exponents ($-1 < \alpha < 0$), the corresponding row section is classified as the well-focused foreground objects. For such a row section, we assign a large depth value to it. This kind of edge points are regarded as the second edge points.

If the Lipschitz exponent of one edge point is large while its adjacent edge point in the same row is small ($-1 < \alpha_1 < 0, 0 < \alpha_2 < 1$ or $-1 < \alpha_2 < 0, 0 < \alpha_1 < 1$), the edge points are regarded as the third kind edge points. If a frame contains lots of the third edge points, it means large amount of row sections between the adjacent edge points are belong to the objects neither in background nor foreground. This frame is lack of distinct stereo feeling and can't be selected as key-frames. Therefore we select the frame with minimum number of the third edge points as a key-frame.

Mallat et al. has demonstrated the Lipschitz exponent of edge point can be calculated by analyzing the evolution across scales of the wavelet transform [18-19]. E_T is the number of the third edge points in a frame of candidate key-frames CF_3 .

The candidate frame with minimum E_T is selected as the optimal key-frame:

$$E_p^f = \frac{E_T^f}{E_A^f}, \quad f \in CF_3 \quad \text{if } E_p^f \text{ is smallest, then } f \text{ is the keyframe} \quad (6)$$



Fig. 3. (a) Color image (b) Frame with occlusion and (c) Depth map

3 Experiment Result

We use two videos to show the performance of our proposed method in experiments. One is the standard video “Kendo”, which has 300 frames with the resolution 1024*768. The other one is a popular American teleplay shot named “Game of Thrones”, which has 300 frames with the resolution 1280*720. An example frame of “Game of Thrones” is shown in Fig. 3(a).

3.1 Objective Evaluation

The depth map generated by the depth estimation algorithm is taken as the original depth. The MSEs between estimated depth maps and propagation depth maps of non-key-frames are calculated and taken as the objective evaluation. The comparison on average MSE is listed in Table 1. 13 frames are selected out as key-frames by different methods and the fixed interval of method in [5] is 25. The depth MSE comparison of each frame is shown in Fig. 4.

For both the “Kendo” and the teleplay shot sequences, the proposed method achieves lower average MSE. Because the left and right view sequences are generated by Depth Imaged Based Rendering, lower depth MSE means that the synthesized 3D video has higher quality. It demonstrates that when a semi-automatic 2D-to-3D conversion is converted to full-automatic one by adopting the proposed method, the quality degradation is the lowest than the other two methods. At the same time, comparing with the full-automatic method which estimates depth map for every frame, computation efficiency of the proposed method is much higher.

Table 1. Average MSE Comparison

Video	Method	Key-frame	Average MSE
Kendo	In [5]	1 25 50 75 100 125 150 175 200 225 250 275 300	314.3600
	In [12]	1 24 45 70 100 120 145 169 193 220 243 264 290	309.3347
	Proposed	1 22 42 79 96 115 136 162 189 214 240 267 293	301.4507
Teleplay	In [5]	1 25 50 75 100 125 150 175 200 225 250 275 300	1200.0
	In [12]	1 30 59 90 114 137 154 175 198 215 237 260 283	1148.1
	Proposed	1 35 63 95 120 146 167 186 206 225 250 267 289	1139.9

3.2 Subjective Evaluation

The main purpose of 2D-to-3D conversion is to synthesize a 3D video with high quality. So the subjective evaluation is an important factor to measure different methods. 20 volunteers, who do not see the experiment results before, participate in our subjective evaluation. The subjective evaluation is shown in Fig. 5. The volunteers watched the video on a 55-inch Samsung 3DTV. The 3D videos converted by different methods are played to each of the 20 volunteers in arbitrary order. For each of volunteers, the time interval between the watching of two different 3D videos should not be less than 2 hours.

Usually the depth and comfort are two trade-off factors. That is to say, distinct depth feeling is easy to cause uncomfortable feeling. However it can be seen from Fig. 5, the proposed method has the highest scores on the five aspects. The proposed method has the best total subjective evaluation scores. It demonstrates that the proposed method can improve the depth and comfort together because the depth of key-frames selected by proposed strategy has high quality. It is a significant advantage of the proposed method. The depth map estimated by wavelet transforming is shown in Fig. 3(c).

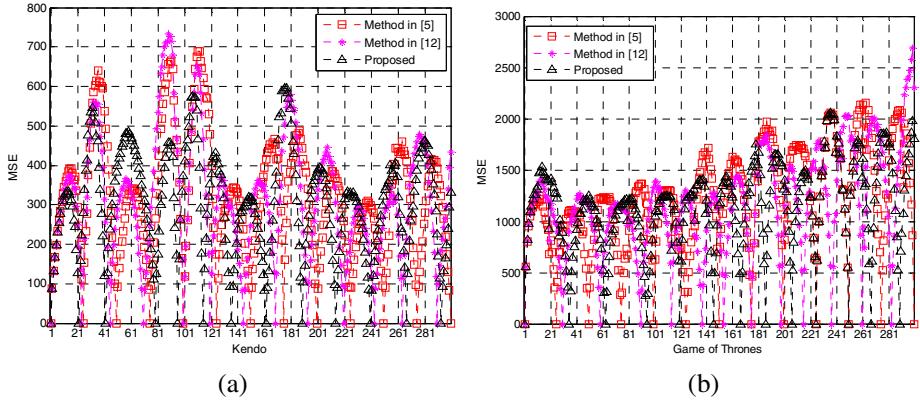


Fig. 4. The MSE Comparison on Different Methods

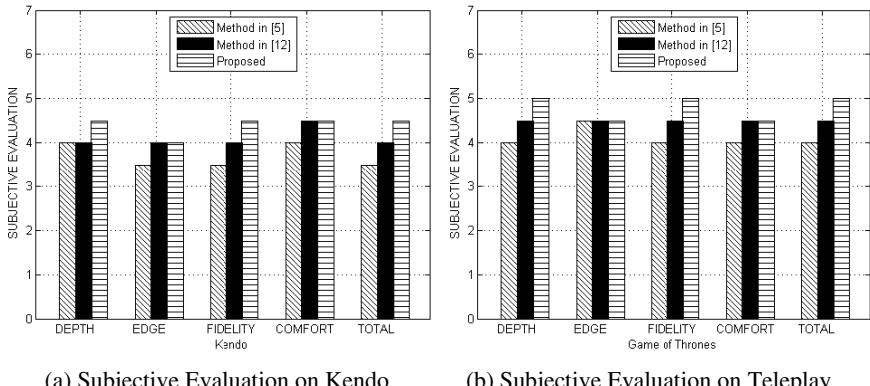


Fig. 5. The Subjective Evaluation

4 Conclusion

A key-frame selection method based on the classification of edge points is proposed in this paper, which is designed for 2D-to-3D conversion. In the proposed method, the optimal key-frames are selected out from the point of depth estimation for the first time. Referring to the depth estimation according to the focus cue, the experiments demonstrate the feasibility and reliability of the proposed method. The performance on the video without ground truth depth maps shows that the proposed method is promising in practical applications. In addition, the semi-automatic 2D-to-3D conversion can be transferred to the full-automatic conversion with little quality degradation by using the proposed method. The optimization on the method of depth estimation according to multi-depth-cue and the evaluation on the quality of the converted 3D video are our future research points.

Acknowledgements. This work was supported in part by the National Basic Research Program of China (No. 2009CB320905), in part by the National Nature Science Foundation of China (No. 61001180 and 61201211), and in part by China Postdoctoral Science Foundation Funded Project (No. 2011M501131). The contact author is Jiande Sun (Email: jd_sun@sdu.edu.cn).

References

1. Varekamp, C., Barebrug, B.: Improved Depth Propagation for 2D-to-3D Video Conversion Using Key-Frames. In: The 4th European Conference on Visual Media Production, pp. 1–7 (2007)
2. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy Snapping. ACM Transaction on Graphics 23, 303–308 (2004)
3. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video Snapcut: Robust Video Object Cutout Using Localized Classifiers. ACM Transaction on Graphics 28, 617–630 (2009)
4. Rotem, E., Wolowelsky, K., Pelz, D.: Automatic Video to Stereoscopic Video Conversion. In: Proceedings of the SPIE, vol. 5664, pp. 198–204 (2005)
5. Cao, X., Li, Z., Dai, Q.H.: Semi-automatic 2D-to-3D Conversion Using Disparity Propagation. IEEE Transaction on Broadcasting 57(2), 491–499 (2011)
6. Gunsel, B., Tekalp, A.M., Peter, J.L.: Moving Visual Representation of Video Objects for Content-based Search and Browsing. In: Proceedings of the ICIP, pp. 502–505 (1997)
7. Zhang, H.J., Wang, Y.A., Altunbasak, Y.: Content Based Video Retrieval and Compression: A Unified Solution. In: Proceedings of the ICIP, pp. 13–16 (1997)
8. Qi, G., Ko, C.C., Sliva, L.C.: A Universal Scheme for Content-based Video Representation and Indexing. In: Proceedings of the IEEE APCCAS, pp. 469–472 (2000)
9. Jia, C., Hou, W.Q., Xu, Y.K.: A New Approach of Key Frame Extraction. Microcomputer Information 21, 133–134 (2007)
10. Nistér, D.: Frame Decimation for Structure and Motion. In: Second European Workshop on 3D Structure from Multiple Images of Large-Scale Environment, pp. 17–34 (2000)
11. Gibson, S., Cook, J., Howard, T., Hubbold, R.: Accurate Camera Calibration for Off-line, Video-based Augmented Reality. In: The 1st International Symposium on Mixed and Augmented Reality, pp. 37–46 (2002)
12. Sun, J.D., Xie, J.C., Liu, J., Shen, Y.J.: Dual Threshold Based Key-frame Selection for 2D-to-3D Conversion. Journal of Computational Information Systems 9(4), 1297–1305 (2013)
13. Valencia, S.A., Rodríguez-Dagnino, R.M.: Synthesizing Stereo 3D Views from Focus Cues in Monoscopic 2D images. In: Proceedings of the SPIE, vol. 5006, pp. 377–388 (2003)
14. Guo, G., Zhang, N., Huo, L.S., Gao, W.: 2D to 3D Conversion Based on Edge Defocus and Segmentation. In: Proceedings of the ICASSP, pp. 2181–2184 (2008)
15. Boykov, Y., Veksler, O., Zabih, R.: Efficient Approximate Energy Minimization via Graph Cuts. IEEE Transactions on PAMI 20(12), 1222–1239 (2001)
16. Kolmogorov, V., Zabih, R.: What Energy Functions can be Minimized via Graph Cuts? IEEE Transactions on PAMI 26(2), 147–159 (2004)
17. Boykov, Y., Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE Transactions on PAMI 26(9), 1124–1137 (2004)
18. Mallat, S., Zhong, S.: Characterization of Signals from Multiscale Edges. IEEE Transactions on PAMI 14, 710–732 (1992)
19. Mallat, S.G., Hwang, W.L.: Singularity Detection and Processing with Wavelet. IEEE Transaction on Information Theory 38(2), 617–643 (1992)

Camera Localization and Pose Estimation Using an RGBD Sensor

Hao Chen and Yan Yuan

Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education,
Beihang University, Beijing, China
jackiechensuper@gmail.com

Abstract. Localization and pose estimation is of great importance for robotic applications e.g. navigation, and mapping. In this paper we developed a localization system based on image data alone. Through feature detection and tracking, we previously build a sparse feature map of an area, then use data captured from an RGBD sensor to track the position and orientation of the camera relative to the map. Matching frames from a live video stream with the underlying map will provide a set of geometric constraints that will allow us to estimate the camera pose. We also compare different methods of calculation, in conjunction with descriptor distance criteria, window region adjustment and RANSAC to decrease the redundancy of the feature point cloud and improve resultant precision. Furthermore, we developed a novel GPU-based implementation for real-time requirement.

Keywords: Localization, Pose estimation, Feature detection, RANSAC, GPU.

1 Introduction

Localization and pose estimation is of great importance for mobile robotic applications e.g. mapping and navigation. A central problem in mobile robotics is how to estimate the 6D pose of a camera relative to a fixed world coordinate system? Building a map of 3D environment whilst localizing the camera is called Simultaneous Localization and Mapping (SLAM) [1]. Pose estimation via registration and SLAM has been studied widely with the help of LiDAR or range sensors [2]. Given the RGB and depth camera parameters, we can reconstruct a point cloud representation of the viewed scene. In indoor environments or areas where GPS/IMU or other measurements are not available or reliable, we must find a new way to have accurate localization and pose estimation.

In this paper we develop a localization system based on image data alone. The system takes a previously acquired dense map of an area as input, constructs a sparse feature map through feature tracking and then uses data captured from an RGBD sensor (e.g. a Kinect camera), to track the position and orientation of the camera.

2 Related Work

Visual odometry refers to the process of estimating an object's 3D motion based only on visual imagery. Moravec was one of the earliest researchers working on this

problem in 1980's [3]. With the development of computer vision, significant progress has been made in the area of visual odometry. For example, many algorithms have been presented for detecting features and describing the local appearance including FAST, SURF, SIFT etc. These feature detectors are fast to compute descriptors and relatively resilient against small viewpoint changes. RANSAC-based methods [4] are also developed for robustly matching features among frames. To enhance visual consistency among frames, Sparse Bundle Adjustment (SBA) [5] can be used to optimize the poses of camera. Sebastian A. Scherer et al. incorporated Bundle Adjustment method in their paper [6]. But there are some inherent problems in PTAM (Parallel Tracking and Mapping) method [7]. With key frames increasing and mapping, the runtime complexity increases as $O(N^3 + N^3M)$. Ivan Dryanovski et al. [8] presented real-time pose estimation by successful registration on successive frames from RGB-D camera. Damian Ancukiewicz et al. [11] suggested using SVD (Singular Value Decomposition) after detecting ASIFT [10] features from a pair of corresponding frames. Recently researchers at Microsoft developed a dense mapping technique [11] known as KinectFusion by using a volumetric representation of a scene, known as a TSDF (Truncated Signed Distance Function), in conjunction with ICP (Iterative Closest Point). This allows rigid pose estimation and alignment of new point cloud data against an existing model.

In Thomas Whelan et al.'s paper [12], they came up with extensions to the KinectFusion algorithm, known as Kintinuous. This method permitted dense mapping of spatially extended regions, incorporating with multiple 6DOF odometry estimation methods for robust tracking. They combine the FOVIS (Fast Odometry from Vision) with both ICP-based and RGBD-based pose estimators. As a result from indoor environmental experiments, the performance demonstrated the ability of Kintinuous system to produce high quality dense color maps with robust tracking in challenging environments.

3 Geometric Model

In geometric projection of pin-hole camera model, there are various coordinate systems: physical world system coordinate $P(X_w, Y_w, Z_w)$, camera coordinate $P(X_c, Y_c, Z_c)$, image pixel coordinate $P(u, v)$, physical image coordinate $P(x, y)$. The following equation can transform image pixel coordinates into world coordinate system for aligning point cloud globally.

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{pmatrix} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} \quad (1)$$

Where f_x, f_y, u_0, v_0 are intrinsic parameters of camera, r_{ij}, t_i are extrinsic parameters. The 6 DOF of camera pose at time i can be represented by a 3×3 rotation matrix R_i and a 3×1 translation matrix T_i . We set the initial pose as $R_0=I$, $T_0=[0, 0, 0]^T$.

4 Camera Localization and Pose Estimation

Given the details discussed above, a proposed schematic for the localization process is illustrated in Fig.1. At the core of the system will be the use of interest points, also known as keypoints or feature points, which provide the means of matching and deriving geometric constraints between the map and the image stream.

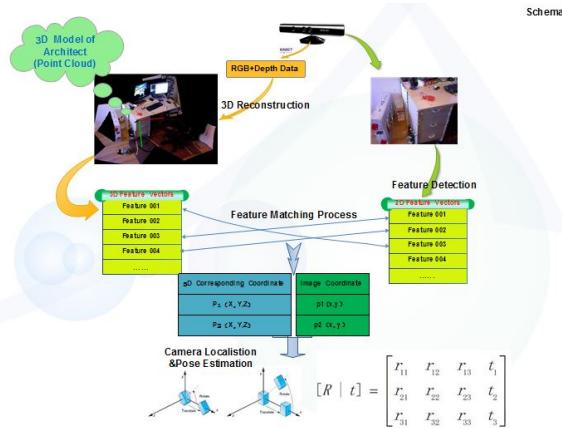


Fig. 1. Project Schematic

4.1 Data Acquisition and Camera Calibration

We can retrieve RGB and depth information by a standard Microsoft camera. The depth and RGB cameras need to be calibrated separately, in order for the intrinsic parameters e.g. focal lengths to be determined. In our work, we calibrated the camera under Z. Zhang's method by nonlinear optimization [14].

4.2 Feature Detection and Tracking

For robustness and computation speed, we choose SURF method for feature detection. Tracking corresponding features precisely has great benefit both in memory and time for 3D model reconstruction. And it does a favor to make pose estimation in real-time possible. By feature matching, we can find the corresponding features, identify and store the new features. In our experiments, we discovered that features often match incorrectly which cause unsatisfactory results by using FLANN-based method. To reject wrong matches, we adopted two methods for precise matching. First, we set criterion on distance between descriptors by finding the minimal distance d_{min} , and set threshold as $T=Kd_{min}$, where $K=3$ in our experiments. We regard values less than T as good matches, otherwise discard them. And we combine window region

adjustment for precise matching. The theory behind this method is that a specific feature only moves within a finite region. So by setting a threshold between its original position and next position, we can reject some bad matches. The tracking result can be seen in Fig. 2 (green lines for trajectory and red points for new features).

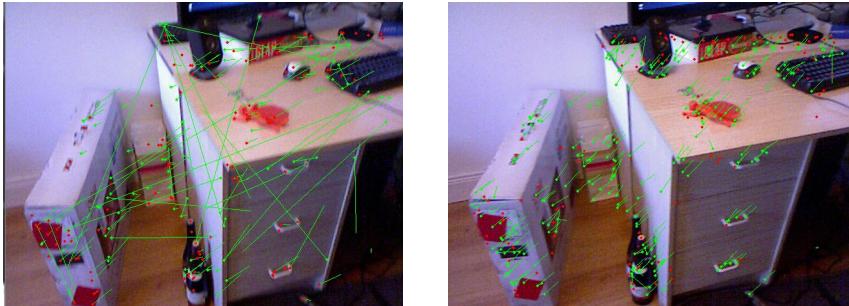


Fig. 2. Tracking results (right figure applied with above methods, while left none)

4.3 Sparse Feature Map Reconstruction

In conjunction with the precise pose estimation which can be achieved in [12], we can project the sparse feature points into a 3D map using following formulas. Through iterating this process we can build a sparse 3D feature map containing both geometric and photometric information.

$$\begin{aligned} \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= R \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + T \\ x &= z * (u - u_0) / f_x \\ y &= z * (v - v_0) / f_y \\ z &= \text{depth} \end{aligned} \tag{2}$$

As the camera moves around an indoor environment, the sparse features are added to the data set incrementally. If we project all the keypoints detected on frames, many points would occupy the same region in reconstruction model. To avoid generating unnecessary redundancy which is negative to computation and visualization, we store each distinct feature descriptor as a new row in a mat called Individual Descriptor Mat (IDM). When a new frame is input, we match the IDM with newly acquired feature. And by matching, we can tell if the feature is distinct. Then add the new features in the IDM as new rows. We call the individual feature building model as Individual Model (IM), while Keypoints Model (KM) by projection all the keypoints. The compared result is quite obvious as seen in Fig. 3. In this case, the IM contains 30564 points while the KM has 64312 points.

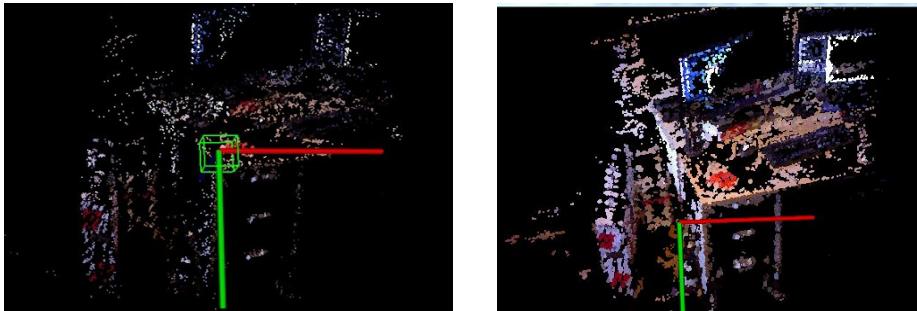


Fig. 3. Individual Model (left) and Keypoints Model (right)

4.4 Match New Features with the Sparse Feature Map

Once the pre-build process is completed, we have the sparse feature map with each data element satisfying two conditions. First, the feature is unique i.e. not same to others. Second, the feature's corresponding depth is a valid value i.e. not infinity or zero. Given an image from camera while the camera is at an arbitrary position and orientation, we provide three methods for matching new acquired features with the available sparse feature map.

- FLANN-based Matcher Method (FM)

FLANN performs fast appropriate nearest neighbor searches in high dimensional spaces. FM trains indexes in the descriptor collection and calls the nearest method to find the best matches.

- Adaptive Distance Criteria Method (ADCM)

The ADCM is similar to the tracking method, but the criteria is much stricter. To avoid the situation where few matches eligible or matches too many, we develop an adaptive method by setting new K value to enlarge or reduce matches for better result.

- Whole Keypoints Method (WKM)

Besides the above two methods, we provide an intact model which stores all the keypoints. This model permits the matching of new feature descriptors with all descriptors to find the best correspondences. We assume this is the best method when no other conditions e.g. computational time, memory storage, are considered. Implementation of camera localization and pose estimation

4.5 Camera Localization and Pose Estimation

After matching new features with the sparse feature map, we can get the n 3D-to-2D point correspondences for localization and pose estimation. In our work, we discuss two methods for this PnP problem.

- Levenberg-Marquardt Iterative Method (LMI).

With Levenberg-Marquardt re-projection optimization in [13], it minimizes the re-projection error i.e. the Sum of Squared Distances (SSD) between the observed image projections and the projected object points. We take 40 frames from the camera at

random positions and then calculate translation errors in three perpendicular axes. For analysis, we take following root-sum-square for error rule as seen in Fig. 4.

$$E_{translation} = \sqrt{X_t^2 + Y_t^2 + Z_t^2} \quad (3)$$

Where X_t , Y_t , Z_t representing translation error in each axe.

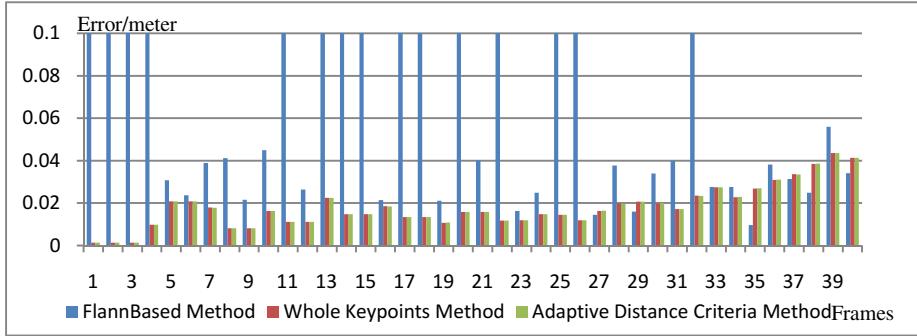


Fig. 4. Translation error using different methods

- Efficient Perspective- n -Point Method (ePnP)

Efficient Perspective- n -Point Method (ePnP) introduces a non-iterative solution with little loss in accuracy but much less computational complexity than regular iterative methods. Its central approach is to take the coordinates of the n 3D points as a weighted sum of four virtual control points detailed in [15].

4.6 Combine Random Sample Consensus (RANSAC) to Reject Outliers

The ADCM method reduces invalid matches but does not achieve adequate levels of precision. To remove the remaining outliers from the collection, we combine RANSAC to reject outlier features. We estimate the pose from a random subset of matches between 2D image points and 3D world points. The resultant pose is used to re-project the remaining 2D points into the 3D world and calculate the re-projection error i.e. SSD. We can find the best subset matches with minimal re-projection error.

5 Experimental Evaluation

5.1 Quantitative and Qualitative Performance for Results

Comparing above three methods that match new features with the sparse feature map, we can see that FM can occasionally calculate wrong results. This was due to the fact that it contained many bad matches which are not even the same corresponding feature. The ADCM costs half of the time compared with WKM while has the similar precise results. For pose estimation, ePnP has little loss in accuracy compared to LMI, while costs less time to compute the pose of the camera. The precision of translation has better result in conjunction with RANSAC method to reject outliers. We calculate

the camera rotation matrix relative to the initial camera center and test it qualitatively in continuous frames by visualization in PCL seen in Fig. 8. Within 300 acquired frames tested, no obvious result occurred.

5.2 Computational Performance

In our system, we used a standard notebook running 32-bit Windows 7 with an Intel(R) Core(TM) 2 Duo CPU T6600 @ 2.2GHz, Nvidia Geforce GT 130M. We compare the time for matching new features with the available sparse feature map on CPU platform in Fig. 5. This step contains detecting the features in a new frame and matching features with the available 3D feature model. We also compare the calculation time of the LMI and ePnP methods in Fig. 6. With the advantage of GPU, we implemented it on GPU for fast optimization and estimation. We achieved the real-time performance at least a rate of 10 Hz. As a result, it can achieve at least a rate of 10 Hz for pose estimation. In Fig. 7, we compare the performance on CPU and GPU platform that GPU can be 15-20 times faster.

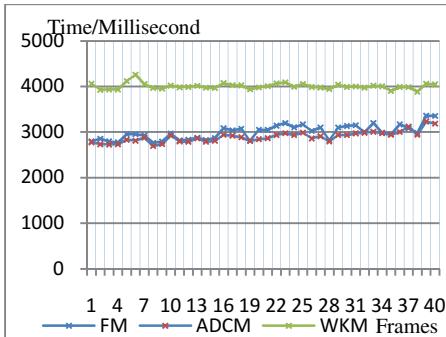


Fig. 5. Detecting, matching and calculation time

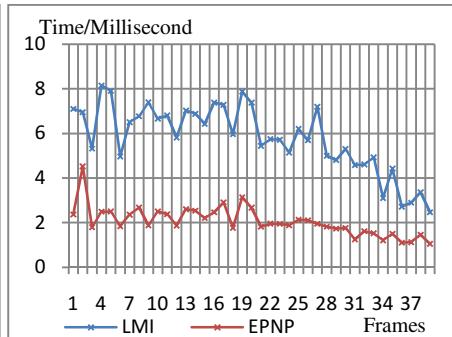


Fig. 6. Calculate pose time (LMI & ePnP)

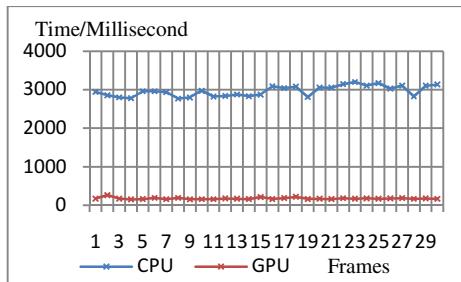


Fig. 7. Comparison of computation time

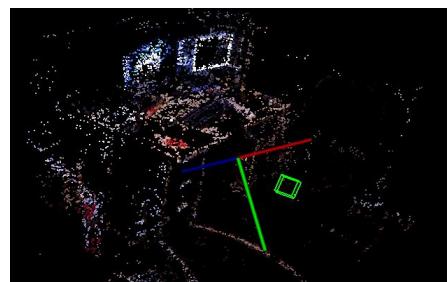


Fig. 8. Visualization of camera

6 Conclusion

In this paper, we developed a RGBD camera localization system based on a previously built sparse feature map. The system accomplishes feature tracking among

consecutive frames by adaptive distance criteria and window region adjustment. This makes the matching process both precise and computationally efficient. Given a set of correspondences between each frame and the sparse feature map after applying RANSAC method, we used them as geometric constraints for estimation. Compared with T. Whelan et al.'s work, our calculation error is not accumulated with time increasing since it's dependent on global pose data, which has been built with robust visual odometry methods.

Acknowledgement. Research presented in this paper was supported by the National Key Basic Research and Development Program (Grant No. 2009CB72400502).

References

1. Hogman, V., et al.: Build a 3D map from RGB-D sensors. Royal Institute of Technology (KTH)
2. Davison, A.J., et al.: MonoSLAM: Real-Time Single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6), 1052–1067 (2007)
3. Moravec, H.: Obstacle avoidance and navigation in the real world by a seeing robot rover (1980)
4. Kaess, M., Ranganathan, A., Dellaert, F.: Incremental smoothing and mapping. *IEEE Trans. on Robotics (TRO)* 24(6), 1365–1378 (2008)
5. Triggs, B., McLauchlan, P.: Bundle adjustment –a modern synthesis. *Vision Algorithms: Theory and Practice*, 153–177 (2000)
6. Scherer, S.A., et al.: Using Depth in Visual Simultaneous Localisation and Mapping. In: 2012 IEEE International Conference on Robotics and Automation, May 14–18 (2012)
7. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2007), Nara, Japan (November 2007)
8. Dryanovski, I., et al.: Real-Time Pose Estimation with RGB-D Camera. In: 2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Hamburg, Germany, September 13–15 (2012)
9. Ancukiewicz, D., et al.: 3D reconstruction from RGB and Depth Video
10. Morel, J.-M., Yu, G.: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences* 2, 438 (2009)
11. Davison, A.J., Izadi, S., et al.: KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality (2011)
12. Whelan, T., et al.: Robust Real-Time Visual Odometry for Dense RGB-D Mapping. In: IEEE Intl. Conf. on Robotics and Automation, ICRA, Karlsruhe, Germany (May 2013)
13. Ranganathan, A.: The Levenberg-Marquardt Algorithm (June 8, 2004)
14. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(11), 1330–1334 (2000)
15. Lepetit, V., et al.: ePnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision* 81, 155–166 (2009)

Sparse Brain anatomical Network Based Classification of Schizophrenia Patients and Healthy Controls

Junjie Zheng¹, Yilun Wang², Heng Chen¹, and Huafu Chen¹

¹ Key Laboratory for NeuroInformation of Ministry of Education,

School of Life Science and Technology,

University of Electronic Science and Technology of China, Chengdu, 610054 China

² The School of Mathematical Sciences,

University of Electronic Science and Technology of China, Chengdu, 611731 China

Abstract. In this study, we tested whether the disturbed structural connectivity in whole brain cortex could be discriminating biomarker for schizophrenia. The anatomical fiber streamlines constructed on AAL template by diffusion tensor image were selected as potential features and a linear SVM pattern classifier was used to categorize the schizophrenia and healthy controls. We randomly divided the whole data into two groups, a training set which contained 32 patients and 25 controls and a test set had 31 patients and 24 controls. We compared two kinds of feature selection methods 1) Univariate t-test based filtering; 2) sparse regression based filtering. The sparse regression features correctly identified 97% cases in test dataset (96% sensitivity and 98% specificity), while the t-test significant impaired connectivity achieved 94% accuracy (92% sensitivity and 96% specificity). The sparse regression selected structural connectivities were consistent in 90% individuals 10 percent more than the t-test filtered features.

Keywords: schizophrenia, structural connectivity, sparse regression, feature selection, classification.

1 Introduction

Through the years, the diagnosis of mental illness is based on clinical manifestation. The more reliable objective measures are needed to help psychiatrists in the process of early stage diagnosis and treatment. Magnetic Resonance Imaging (MRI), which is the non-invasive investigation of the structure and function pathology of the human brain has been put in use for classification the clinical psychiatric disease. A growing numbers of studies extracted more information for neuroimaging data using machine learning methods to distinguish the mental disorder for controls or to find potential neurological biomarker [1]. As in schizophrenia, many researchers have shown the advantages of functional or structural MRI and diffusion tensor image in classification [2,3].

Schizophrenia is believed to result from abnormal functional integration of neural processes due to the alteration of interactions between two or more regions. The evidence of widespread impaired connectivity in white-matter connectional architecture in schizophrenia has been presented [4,5]. Whether these disturbed connects which

presented as the brain network edge and characterized the brain network small world-ness could be the discriminating biomarkers largely remains unknown.

In this study, we used diffusion tensor image to provide a measure of brain anatomical connectivity through white matter tracts. We extracted the strength of the network edges measured by number of cortical-cortical streamlines as the potential features and preformed SVM classification of the schizophrenia and healthy controls after two different feature selections for the mass 4005 edges. As the feature selection is needed to obtain optimal accuracy especially in high dimensional data [6], we compared the total classify accuracy and stability of univariate t-test based filtering and sparse regression based filtering method for brain network edge selection.

2 Materials and Methods

2.1 Subjects

Forty-nine antipsychotic-naive patients with first episode schizophrenia and were Sixty-three individuals without psychiatric diagnoses matched by age, sex, and years of education were both recruited from Hunan province through the outpatient Psychiatric Department of the Second Xiangya Hospital of Central South University.

The diagnosis of schizophrenia illnesses was performed using the Structured Clinical Interview from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (SCID) [7]. All the patients were first-episode, drug-naive schizophrenics diagnosed formally according to the DSM-IV-TR schizophrenia. Patients were also evaluated using the Positive and Negative Symptom Scale (PANSS) [8] by a senior psychiatrist.

2.2 Image Processing

Brain MRIs were acquired with a 3.0 Tesla Philips scanner (Philips Medical Systems). The diffusion tensor images covering the whole brain were obtained using spin echo-based echo planar imaging sequence, including 33 volumes with diffusion gradients applied along 33 non-collinear directions ($b = 1000 \text{ s/mm}^2$) and one volume without diffusion weighting ($b = 0 \text{ s/mm}^2$). High-resolution T1-weighted images of the whole brain were acquired in a sagittal orientation with the following parameters: repetition time = 7.4 ms, echo time = 3.4 ms, inversion time = 875 ms, flip angle=9°, acquisition matrix=228×228, field of view=250 mm×250 mm, slice thickness=1.1 mm gap=0 mm and slices=301.

For each subject, diffusion-weighted images were geometrically corrected using a non-diffusion-weighted B0 image ($b = 0 \text{ s/mm}^2$) and a field map, to account for eddy current distortions. Diffusion-weighted images were co-registered to the B0 image using affine transformations to minimize slight head movements. Diffusion tensor models were estimated by the linear least-squares fitting method at each voxel using the Diffusion Toolkit (trackvis.org) [9]. Whole-brain fiber tracking was performed in native diffusion space for each subject using the Fiber Assignment by Continuous Tracking (FACT) algorithm embedded in the Diffusion Toolkit. Path tracing proceeded until either the fractional anisotropy was less than 0.1 or the angle between the current

and the previous path segment exceeded 45 degree. To determine the nodes of structural connectivity network in each subject, AAL regions of interest (ROI) were defined in native diffusion space. This procedure has been applied in previous study [10]. The two ROIs were considered to be connected through an edge, in case at least one fiber were presented between them. For each edge, we calculated the tracked fiber streamlines as the connectivity strength between two regions.

2.3 Feature Selection

Cortical-cortical and cortical-subcortical fiber track connects were constructed for each person and were modeled as the network or graph with 90 nodes. We then extracted each paired regions connected streamlines resulting $90 * (90 - 1)/2 = 4005$ connects as features for next machine learning process to train and classify. For some features are less sensitive, irrelevant or redundant for the classify performance [6], we filtered the these features by two methods. We first ranked the brain network connectivity by the two group comparison t-test P values duo to the most significant differently disturbed ones having more potential for discrimination [6]. We cut the widely used statistic threshold $p < 0.05$ for selection and returned N connectivity features from anatomical brain network.

Recently, the sparse regression based feature selection methods have been widely used and developed for ensuring efficient use of data and faster computation time for small samples comparing mass features. SVM based feature selection using L1, L2 or L0 regularization methods and logistic regression with a combination of L1, L2 norm regularization for accurately discrimination were also proposed in the literatures [11,12]. Here, we formulated matrix of the training set with each row is a 4005 dimensional feature-vector from one subject and the labels 1 for healthy and -1 for patients subjects. The sparse regression as below (1) returned a coefficient vector and the absolute of weight for each feature indicated the contribution of the corresponding feature to discriminating the two groups. We ranked them by the weights and also selected N features as the t-test filtering preventing the bias of two selections.

$$\min_{x \geq 0} \frac{1}{2} \|y - Ax\|^2 + \lambda \|x\|_1 \quad (1)$$

We solved the L1-norm regularized least squares problem where y represents sample labels 1 and -1, and A is the features matrix with the row data indicate samples and column data are features. We chose $\lambda = 0.2$ for the formula and got x as a vector represented the weights for features. We also tried other values of parameter λ , and the results were similar.

2.4 Discriminant Analysis and Performance Estimation

In this study, we randomly divided the whole schizophrenia and control samples into two groups. One group contained 32 patients and 25 controls as training set and there were 31 patients and 24 healthy control samples in the other set as test set. Support

vector machine was used to solve the classify problem of separate the schizophrenia patients from healthy controls. We trained the first group with selected features above and validated the performance in the test set. The final quality of a model was assessed by Sensitivity=TP/(TP+FP), Specificity=TN/(TN+FN) and Accuracy=(TP+TN)/(TP+FP+TN+FN), where TP is the number of true positives (correctly classified patients), and FP is it the number of false positives while TN is the number of true negatives (correctly classified control samples), and FN is the number of false negatives. The whole procedure was figured out as Fig.1.

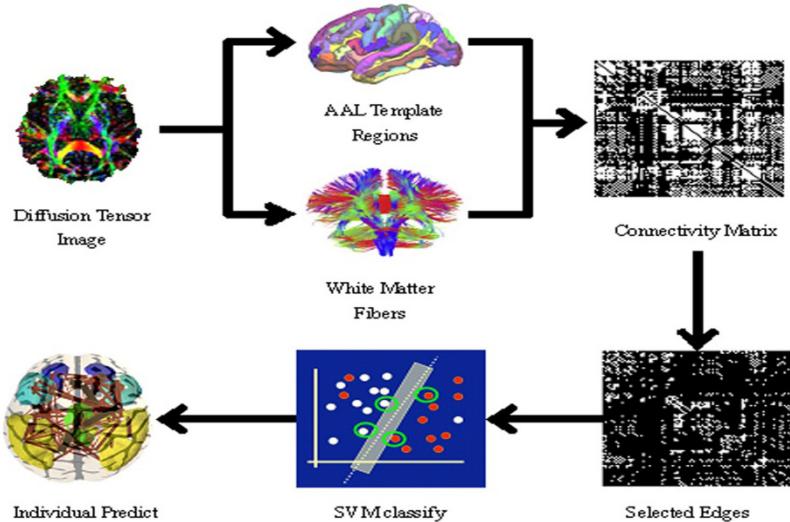


Fig. 1. Flow diagram illustrating the whole image data procession, brain network construction, connectivity feature selection and SVM classify procedural

We then tested the consistency of the selected discriminative features for the high variance of connectivity streamlines across subjects [13]. To measure this consistency, we defined the subject-specific binary quantity C , such that $C_{ij} = 1$ for each feature i if connectivity between two cortical regions had one or more white matter tracts within a subject j , and $C_{ij} = 0$ otherwise as [13]. The consistency of connectivity i was calculated by $\text{Cons}_i = \sum_1^n C_j / n$ where n is the number of samples. A group of potential features had the mean consistency of them. Then we compared the top 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100% selected group features meanCons between the t-test features filtering and sparse regression selection.

For stably classification, we randomly did 100 permutations of the group division. In each time of repetition, we randomly pick 32 patients and 25 controls as train set from the whole samples of which the remains composed the validation data. For each random group parcellation, we estimated the classification performance accuracy and features consistency comparison. The whole process was performed in MatlabR2011a integrating the LIBLINEAR [14] and our software.

3 Results

For each random group parcellation, we selected about 458(398-562) significantly changed structural connectivity by t-test between schizophrenia and healthy control samples anatomical network. We also filtered the same number high weighted features from train dataset by sparse regression. Then the SVM trainings were performed on both feature sets. The uni-variate t-test filtering features resulted about 92% (75%-100%) sensitivity, 96%(87%-100%) specificity, and 94%(87%-100%)total accuracy while the sparse regression based connectivity showed 96%(87%-100%) sensitivity, 98%(87%-100%) specificity, and 97%(89%-100%)total accuracy as Fig2. The improvements of sensitivity, specificity and accuracy by sparse regression feature selection are significant and the performance stability form permutation was better than t-test feature filtering.

The features consistency comparison between the two selection methods showed that t-test based features had 78% to 90% consistency in train set and 80% to 97% consistency in test set where sparse regression high weighted features had 90% to 98% consistency in both train set and test set. The results also indicated that the more high weighted features selected by sparse regression had more consistency in individual brain network but the significantly changed connectivity from uni-variate statistic test. See Fig2.

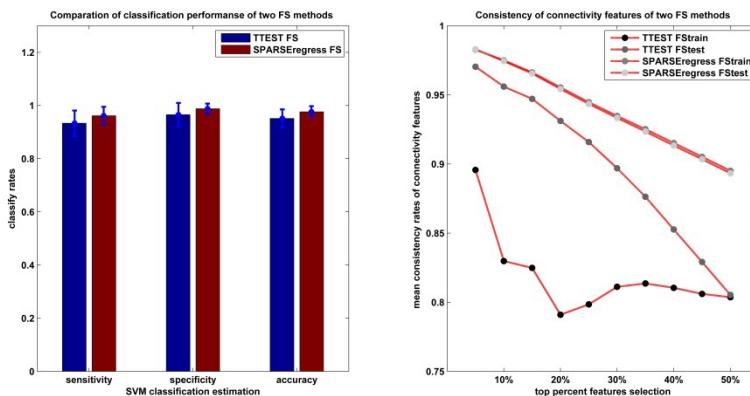


Fig. 2. The left side bar plot indicated t-test and sparse regression based feature selection classify performance on 100 randomly divided data sets; The right lines showed the connectivity features consistency from the two filtering methods in train set and test set. FS: feature selection.

4 Discussion and Conclusion

We extracted the paired cortical to cortical or cortical to sub cortical regions connectivity from anatomical network as the potential features, and used the support vector machine to classify the schizophrenia and healthy controls. The high

classification performance in sensitivity, specificity and total accuracy indicated that the brain structural network indeed contained the discriminating neurological biomarkers for schizophrenia. These widely disturbed fiber streamlines detected by tractography across the whole brain do not only disconnect but affect the whole brain net neural information integration and transportation. In this study we showed that how the foundational connectivity of disturbed Small Worldness in schizophrenia were taken together to predict individual brain network abnormality [5].

We compared the two classical feature selection methods before classification especially the mass dimensional characters while small sample set. The univariate statistic t-test selected significantly impaired connectivity from brain structural network, but this method considered the independent effects of the network wirings and would come out more relevant false positive edges when we research on large scale brain network. We performed the sparse regression as feature selection which had more high accuracy. Additionally, it resulted the stable classification and more consistent brain impaired connectivity. We showed the sparse regression method advantages in dealing with 90 paired regions brain network and we would benefit more when meeting the large scale network with huge pathways or circuits in it.

Acknowledgments. This work was supported by the Natural Science Foundation of China, Grant Nos. 61035006, 61125304 and by the Specialized Research Fund for the Doctoral Program of Higher Education of China 20120185110028. This work also was supported by the Natural Science Foundation of China, Grant No. 11201054 and by the Fundamental Research Funds for the Central Universities ZYGX2012J118.

References

1. Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A.: Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews* 36, 1140–1152 (2012)
2. Ardekani, B.A., Tabesh, A., Sevy, S., Robinson, D.G., Bilder, R.M., et al.: Diffusion tensor imaging reliably differentiates patients with schizophrenia from healthy volunteers. *Human Brain Mapping* 32, 1–9 (2011)
3. Nieuwenhuis, M., van Haren, N.E., Hulshoff Pol, H.E., Cahn, W., Kahn, R.S., et al.: Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 61, 606–612 (2012)
4. Fornito, A., Zalesky, A., Pantelis, C., Bullmore, E.T.: Schizophrenia, neuroimaging and connectomics. *Neuroimage* (2012)
5. Zalesky, A., Fornito, A., Seal, M.L., Cocchi, L., Westin, C.-F., et al.: Disrupted axonal fiber connectivity in schizophrenia. *Biological Psychiatry* 69, 80–89 (2011)
6. Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C., et al.: Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage* 60, 59–70 (2012)
7. First, M.B., Gibbon, M.: User's guide for the structured clinical interview for DSM-IV axis I disorders: SCID-1 clinician version: American Psychiatric Pub. (1997)
8. Kay, S.R., Fiszbein, A., Opler, L.: The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* 13, 261–276 (1997)

9. Wang, R., Benner, T., Sorensen, A., Wedeen, V.: Diffusion toolkit: a software package for diffusion imaging data processing and tractography, 3720 (2007)
10. Zhang, Z., Liao, W., Chen, H., Mantini, D., Ding, J.-R., et al.: Altered functional-structural coupling of large-scale brain networks in idiopathic generalized epilepsy. *Brain* 134, 2912–2928 (2011)
11. Ryali, S., Supekar, K., Abrams, D.A., Menon, V.: Sparse logistic regression for whole brain classification of fMRI data. *NeuroImage* 51, 752 (2010)
12. Bi, J., Bennett, K., Embrechts, M., Breneman, C., Song, M.: Dimensionality reduction via sparse support vector machines. *The Journal of Machine Learning Research* 3, 1229–1243 (2003)
13. Hermundstad, A.M., Bassett, D.S., Brown, K.S., Aminoff, E.M., Clewett, D., et al.: Structural foundations of resting-state and task-based functional connectivity in the human brain. *Proceedings of the National Academy of Sciences* 110, 6169–6174 (2013)
14. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9, 1871–1874 (2008)

Sparse Learning for Face Recognition with Social Context

Jie Gui^{1,2,3}, Jian-Xun Mi⁴, Ying-Ke Lei⁵, and Hong-Qiang Wang²

¹ State Key Laboratory of Software Engineering, Wuhan University, 430072

² Hefei Institute of Intelligent Machines, Chinese Academy of Sciences,
Hefei, Anhui, China

³ State Key Laboratory for Novel Software Technology, Nanjing University, P.R. China

⁴ Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology,
Shenzhen, Guangdong Province, China

⁵ Electronic Engineering Institute, Hefei, Anhui 230037, China

Abstract. Face recognition in uncontrolled environments, such as large pose variations, and extreme ambient illumination and expressions is a challenging task for traditional face recognition methods. Some recent works show that context information such as clothes and social relationships is very important for solving the problem. Furthermore, sparse representation-based method is very robust for face occlusion and pixel contamination. In this paper, the authors propose a sparse learning framework for face recognition with social context. First, sparse representation based dimensionality reduction method is used to find the low dimension representation of the face images. Second, sparse representation based classification is utilized to classify test face images as known label and unknown label. Third, for test images classified as unknown label, social context descriptor is constructed according to co-existence information. Finally, based on social context descriptor and the low dimension representation of the test images classified as unknown label, sparse representation based clustering is adopted to perform face recognition.

Keywords: Face recognition, social context, sparse representation based classification, sparse representation based clustering, sparse representation based dimensionality reduction.

1 Introduction

The aim of face recognition (FR) is identifying or verifying a face from its image. It has received great attention over the last decades due to its value both in understanding how FR process works in humans and its real-world applications, such as access control and video surveillance.

Traditional FR methods train a classifier by using labeled face images and then recognize unlabeled images through utilizing the classifier. Up to now, an enormous volume of literature has been devoted to develop this kind of methods. These FR methods have already achieved impressive performance in controlled conditions.

However, in real applications, such as large pose variations, and extreme ambient illumination and expression variations (Fig. 1, Fig. 2 and Fig. 3), this is a challenging task for traditional face recognition methods.



Fig. 1. Some typical samples of the cropped images with different poses in FacePix face database



Fig. 2. Some typical samples of the cropped images with different illuminations in Extended Yale B face database



Fig. 3. Seven typical samples of the cropped images with different expressions in AR face database

Some recent works show that context information [1-4] such as clothes and social relationships is very important for solving the problem. There are two figures Fig. 4 and Fig. 5 to illustrate this problem. In the right figure of Fig. 4, there is a girl in the left corner of this figure. Although it is difficult to see the face details of the girl, it is easy to infer that there was a high probability the girl is Obama's daughter according to the co-existence information. In Fig. 5, it is very difficult even for persons to determine how many girls are shown and which images are of the same girls from only the girls (top). However, when the faces are embedded in the context of clothing, it is much easier to recognize the three girls (bottom).



Fig. 4. The president of the United States Obama's family



Fig. 5. Three girls ([4])

Recently, some new machine learning methods integrating the theory of sparse representation, compressed sensing have been proposed, and have been successfully applied in face recognition [5-7]. Sparse representation based classification (SRC) [5] demonstrates significant improvements in accuracy over traditional face-recognition techniques. SRC has been reported in the journal “Communications of the ACM” with the title of “face recognition breakthrough”. Sparse representation based clustering (SC) has been proposed in [8]. Sparse subspace learning (SSL) [9] is a special family of dimensionality reduction methods which consider “sparsity”. It has either of the following two characteristics: (1) Finding a subspace spanned by sparse base vectors. The sparsity is enforced on the projection vectors and associated with the feature dimension. The representative methods include sparse principal component analysis (SPCA) [10], sparse nonnegative matrix factorization [6], and nonnegative sparse PCA [11], etc. (2) Aiming at the sparse reconstructive weight which is associated with the sample size. The representative methods include sparse neighborhood preserving embedding (SNPE) [8]. In fact, SNPE is identical to sparsity preserving projections (SPP) [12], which has achieved higher recognition rates than PCA and neighborhood preserving embedding (NPE) for face recognition. Zhang et al. [13] proposed a sparse representation-based classifier (SRC) [5] oriented unsupervised dimensionality reduction algorithm which combines SRC and PCA in its objective function. Yang et al. [14] proposed the SRC steered discriminative projection (SRC-DP). The basic idea of

SRC-DP is to seek a linear transformation such that in the transformed low-dimensional space, the within-class reconstruction residual is as small as possible and simultaneously the between-class reconstruction residual is as large as possible.

In this paper, the authors propose a sparse learning framework for face recognition with social context. Algorithm model will be presented in Section 2.

2 Algorithm Model

For natural images with faces, we will perform face detection, face segmentation and face alignment to obtain the normalized face images.

Since the dimensionality of face images is usually very high, we first perform sparse dimensionality reduction (SDR) to get the low dimensionality representation of all face images.

The key step of a pattern recognition system is classification. After obtaining the low dimensionality representation of all face images, sparse representation based classification (SRC) will be used as the classifier. Each test face image will be classified as known class or unknown class by using sparsity concentration index (SCI) [5]. SCI is defined as

$$SCI(x) = \frac{k \cdot \max_i \left\| \delta_i(x) \right\|_1 / \left\| x \right\|_1 - 1}{k - 1} \in [0, 1]$$

where k denotes the number of classes, x denotes the test sample and $\delta_i(x)$ is solved by SRC. For a solution \hat{x} found by SRC, if $SCI(\hat{x}) = 1$, the test image is represented using only images from a single object, and if $SCI(\hat{x}) = 0$, the sparse coefficients are spread evenly over all classes. We choose a threshold $\tau \in (0, 1)$ and classify a test image as known classes if $SCI(\hat{x}) \geq \tau$ and otherwise classify it as unknown classes.

For test image classified as unknown classes, we will construct social context descriptor (SCD) to reflect his co-existence information. Since there are k classes, each test image will have k residuals. The residuals of the j^{th} sample is denoted by $r_{ji}(x), i = 1, \dots, k$. The smaller the residual is, the more similar the test image is to the corresponding class. Suppose m persons co-exist in a image, the SCD is defined

as $\sum_{j=1}^m \left[\frac{1}{r_{j1}(x)}, \frac{1}{r_{j2}(x)}, \dots, \frac{1}{r_{jk}(x)} \right]$. Furthermore, we can integrate the age, the background of the photo and the description information of the photo in SCD.

Finally, the SCD and LDR of all unknown class test images are combined to form a long vector. Based on these long vectors, the similarities of faces are computed and sparse representation based clustering is adopted to perform face recognition.

So to sum up, the algorithm model is summarized in the following:

Step 1. Performing sparse dimensionality reduction (SDR) to get the low dimensionality representation of all face images;

Step 2. Performing sparse representation based classification to classify all unlabeled face images;

Step 3. Constructing social context descriptor (SCD);

Step 4. Performing sparse representation based clustering (SC).

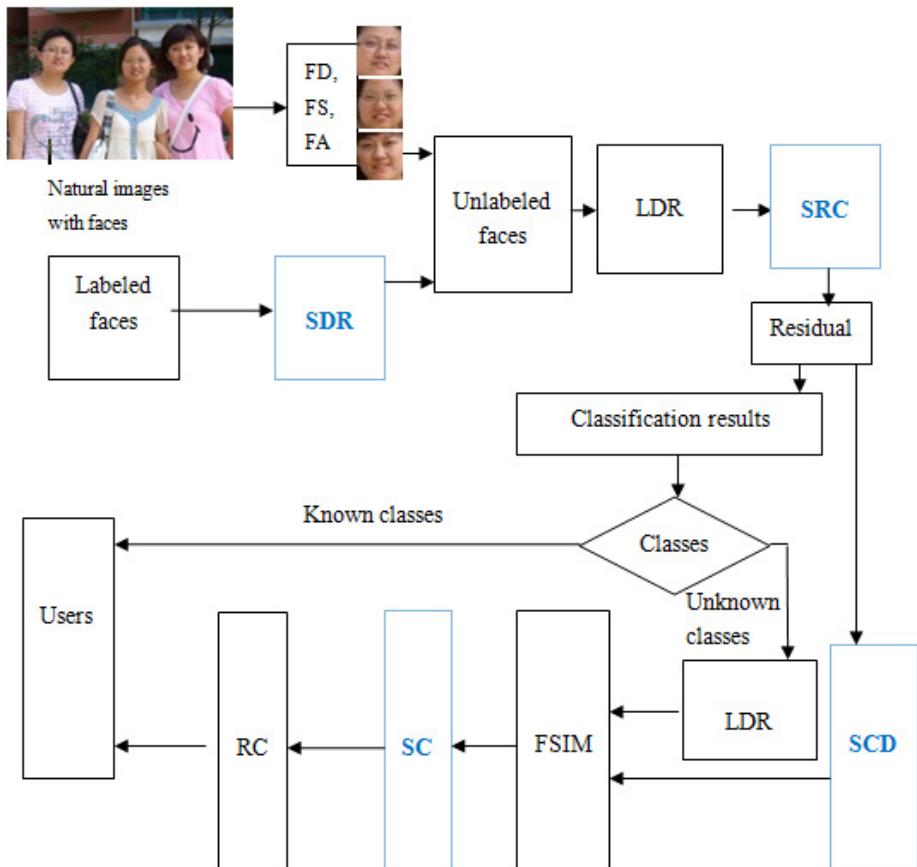


Fig. 6. Sparse learning framework for face recognition with social context

Since there are many abbreviations in Fig. 6, we summarize these abbreviations in Table 1 to make Fig. 6 clear.

Table 1. Abbreviation vs Full name

Abbreviation	Full name
FD	Face detection
FS	Face segmentation
FSIM	The similarity between faces
FA	Face alignment
SDR	sparse dimensionality reduction
SRC	sparse representation based classification
SC	sparse representation based clustering
RC	results of clustering
LDR	Low dimensionality representation
SCD	Social context descriptor

3 Conclusions

In this paper, the authors propose a sparse learning framework for face recognition with social context. We plan to construct a face database with social context and perform corresponding experiments to verify our algorithms in our future work. Note that this is a flexible framework. Any SDR can be used. For specific face recognition application, choosing which SDR remains an open problem. SCD can be defined to integrate more information. This deserves further studies.

Acknowledgement. This work was supported by the National Science Foundation of China (Grant No. 61100161, 61175022, 61005007, 61272333, 61005010, 31271412), the Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. Y023A61121, Y023A11292), Open Fund Project of State Key Laboratory of Software Novel Technology, Nanjing University, P.R. China (KFKT2012B26), Open Fund Project of State Key Laboratory of Software Engineering, Wuhan University, China (SKLSE2012-09-25), Key Project of Natural Science Research of Anhui Provincial Education Department, No. KJ2013A076 and Anhui Provincial Natural Science Foundation (grant no. 1208085MF96).

References

- [1] Singla, P., Kautz, H., Luo, J., Gallagher, A.: Discovery of Social Relationships in Consumer Photo Collections using Markov Logic. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 1-3, pp. 735–741 (2008)
- [2] Stone, Z., Zickler, T., Darrell, T.: Toward Large-Scale Face Recognition Using Social Network Context. Proceedings of the IEEE 98, 1408–1415 (2010)
- [3] Lin, D., Kapoor, A., Hua, G., Baker, S.: Joint People, Event, and Location Recognition in Personal Photo Collections Using Cross-Domain Context. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 243–256. Springer, Heidelberg (2010)

- [4] Gallagher, A.C., Chen, T.: Using Context to Recognize People in Consumer Images. *IPSJ Transactions on Computer Vision and Applications* 1, 115–126 (2009)
- [5] 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, 210–227 (2009)
- [6] Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* 5, 1457–1469 (2004)
- [7] Pu, X., Yi, Z., Zheng, Z., Zhou, W., Ye, M.: Face recognition using fisher non-negative matrix factorization with sparseness constraints. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) *ISNN 2005. LNCS*, vol. 3497, pp. 112–117. Springer, Heidelberg (2005)
- [8] Cheng, B., Yang, J.C., Yan, S.C., Fu, Y., Huang, T.S.: Learning With $\ell(1)$ -Graph for Image Analysis. *IEEE Transactions on Image Processing* 19, 858–866 (2010)
- [9] Cai, D., He, X., Han, J.: Spectral regression: A unified approach for sparse subspace learning. Presented at the International Conference on Data Mining (ICDM 2007), Omaha, NE (2007)
- [10] Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15, 265–286 (2006)
- [11] Zass, R., Shashua, A.: Nonnegative sparse PCA. *Advances in Neural Information Processing Systems* 19, 1561 (2007)
- [12] Qiao, L.S., Chen, S.C., Tan, X.Y.: Sparsity preserving projections with applications to face recognition. *Pattern Recognition* 43, 331–341 (2010)
- [13] Zhang, L., Yang, M., Feng, Z., Zhang, D.: On the Dimensionality Reduction for Sparse Representation based Face Recognition. In: 2010 International Conference on Pattern Recognition (2010)
- [14] Yang, J., Chu, D.: Sparse Representation Classifier Steered Discriminative Projection. In: 2010 International Conference on Pattern Recognition (2010)

Finger Vein Recognition Based on Gabor Filter

Hong Zhang¹, Zhi Liu^{1,*}, Qijun Zhao², Congcong Zhang¹, and Dandan Fan¹

¹ School of Information Science and Engineering,
Shandong University, Jinan, 250100, P.R.C.
liuzhi@sdu.edu.cn

² College of Computer Science, Sichuan University,
Chengdu, 610065, P.R.C.

Abstract. Finger vein recognition is a promising biometric authentication technique. Finger vein images include a plurality of lines and can be regarded as a type of texture image. This paper proposes the use of 2D Gabor filters to process finger vein images and extract the texture features for better recognition results. Euclidean distance matching is performed. Experimental results demonstrate the effectiveness of this method.

Keywords: Finger vein recognition, 2D-Gabor filter, Feature extraction.

1 Introduction

Many people devoted themselves to biometrics-based identification research in the past several decades. A perfect biometrics-based identification system is difficult to establish although identification can be based on various types of biometric features, such as the face, iris, and fingerprint. Finger vein recognition technology as a newly developed biometric technology has the following advantages: immunity to touch, interior feature, stability, and uniqueness. Hence, finger vein recognition technology has attracted increasing attention. Although finger vein recognition technology has been initially used to access control and security systems in companies, schools, and other organizations, some difficulties still need to be addressed for this biometric pattern to be widely used.

The main difficulty in finger vein recognition technology is extracting the finger vein texture because of the poor quality of captured finger vein images [1]. Many researchers around the world established methods to enhance the finger vein image or extract the vein texture [2]. Miura et al. [3] in 2007 proposed the method of utilizing maximum curvature points in image profiles to obtain finger vein patterns. Yu Chengbo et al. [4] proposed the method of finger vein texture extraction based on the region growth and the method based on detection of valley lines [5]. An algorithm first extracts vein texture approximately by detecting valley lines and then obtains the texture by segmenting image threshold and enhancing blur. Li Hongbing et al. [6] explored ridgelet transformation to enhance finger vein images. Yang

* Corresponding author.

Jinfeng et al. [7] employed filter banks to avoid down sampling direction and Frangi filtering to enhance finger vein images. Yang Xin et al. [8] proposed the method for feature extraction with sparse representation. The abovementioned methods have a certain optimization effect on feature enhancement and extraction in low-quality finger vein images. The low contrast of finger vein images, large amount of noise, and lack of periodic texture are problems that have an impact on the final recognition results. This paper proposes a finger vein recognition algorithm based on Gabor filters to resolve these problems.

2 Acquisition and Normalization of the Finger Vein Image

Near-infrared light is utilized in finger vein recognition to illuminate the fingers. A ray with a specific wavelength penetrates the skeleton and muscle of a finger; at the same time, hemoglobin absorbs the light radiation efficiently. An illustration of the near-infrared light camera that is utilized for finger vein image acquisition is shown in Fig. 1. The captured finger vein images are shown in Fig. 2.

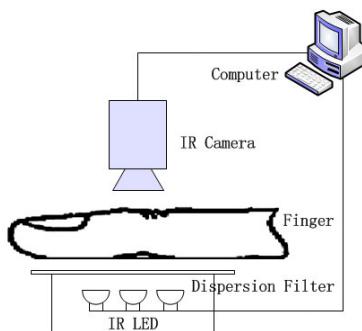


Fig. 1. Illustration of the developed imaging device

The top point of the finger is first fixed, and then the middle part of the finger is extracted by measuring the middle three-fifth of the finger. The two points obtained from the upper line crossing the counter of the finger are then utilized to set a rectangle to obtain the region of interest (ROI). Normalizing the size of the processed image is necessary because the ROI of the finger vein image has different sizes among individual fingers. The results of this procedure are shown in Fig. 3.

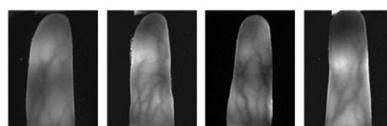


Fig. 2. Captured finger vein images

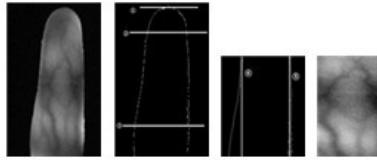


Fig. 3. Captured finger vein images

3 Finger Vein Recognition Based on Gabor Filter

3.1 Theory of Gabor Filter

The 2D Gabor filter is a type of band-pass filter that has directional and frequency selectivity and achieves the best combination of the spatial and frequency domains. A spatial even-symmetric Gabor filter is generally defined as

$$G(x, y : f, \theta) = \exp\left\{-\frac{1}{2}\left[\frac{x_\theta^2}{\sigma_x^2} + \frac{y_\theta^2}{\sigma_y^2}\right]\right\} \cos(2\pi f x_\theta) \quad (1)$$

θ is the orientation of the Gabor filter, f denotes the filter center frequency, σ_x and σ_y represent the space constants of the Gaussian envelope along the coordinate axes x and y , respectively. These parameters determine the bandwidth of the band-pass filter. Equation (1) shows that a 2D even Gabor filter is a function composed of a 2D Gaussian-shaped function and cosine function.

The features of the Gabor filter are very similar to the human visual system, particularly in representing frequency and direction. Therefore, the 2D Gabor filter is often utilized to extract texture features. Many scholars have achieved very satisfactory results by applying Gabor filters in fingerprint and iris recognition [9,10]. Considering the lack of singularities in the local neighborhood, finger vein images are constituted by valley and ridge lines, which present uniform thickness. Figure 4 shows that valley and ridge lines are curvy, continuous, and vary smoothly in direction in most areas [2]. Therefore, the finger vein pattern can be regarded as a kind of texture image.

Owing to its spatial characteristics, the 2D Gabor filter can maximize the local direction and frequency information of an image. After filtering the finger vein image through the 2D Gabor filter, we can obtain the enhanced ridge structure of the finger vein. The ridge structure has a specific spatial frequency and direction.



Fig. 4. Texture of a finger vein image

3.2 Significance of the Parameters of the Gabor Filter

Considering the direction selectivity of the Gabor filter, filters are utilized in different directions to filter finger vein images. Half of these directions may be ignored because selection by the Gabor filter is even symmetric; the even symmetry of the filter utilized for the image and its response remain unchanged in the phase rotation. This study adopts m directions $[0, \pi]$ as follows,

$$\theta = \frac{k}{m} * \pi k = 0, 1, ; \dots, (m - 1) \quad (2)$$

Employing a Gabor filter on the image in each direction yields an enhanced image in a specific direction. An increase in the value of m provides more filtered images, which in turn provides more extracted features. This relationship is helpful for identification. However, the filter will reduce the noise tolerance of images if m is too large. The value of m in this study is 8. Therefore, the filter can be obtained in eight directions: $0^\circ, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ, 112.5^\circ, 135^\circ, 157.5^\circ$.

Eight enhanced images can be obtained after filtering the finger vein image with eight filters. We can extract a feature vector from each of the eight images.

f represents the central frequency of the filter. f is denoted by the characteristics of the signal: when the texture frequency of the filter and texture of the image frequency of the filter are similar and when the amplitude of the signal after filtering is at maximum. A 2D Gabor filter has frequency selectivity; thus, information can be extracted by adjusting the value of f to the desired specific frequency. False lines are produced in 2D Gabor filtering when the value of f is too large. In contrast, the image becomes blurred when the value of f is too small.

σ_x and σ_y denote the space constant of the Gaussian envelope along the x and y axes, respectively, and determine the bandwidth of the band-pass filter. The image generates a pseudo ridge line when the bandwidth is too large. Noise elimination is reduced when the bandwidth is too small.

3.3 Feature Extraction and Comparison of Feature Values

We extract vein features from the eight Gabor-filtered images. The standard deviation in the window with an $r \times r$ size is obtained by Equation (3). The obtained standard deviation is regarded as the feature values of the region.

$$s(x, y) = \sqrt{\frac{1}{r^2} \sum_{i=x-r/2}^{x+r/2} \sum_{j=y-r/2}^{y+r/2} ((f(i, j))^2 - (m(x, y))^2)} \quad (3)$$

Assuming that the size of the normalized image is $(m \times m)$, the entire image is divided into $(m \times m)/(r \times r)$ blocks. Therefore, the image is extracted from $M = (m \times m)/(r \times r)$ feature values. We obtain images in eight directions after filtering, resulting in a total of $8 \times M$ feature values.

The similarity of the images is examined after feature extraction to determine if the images are from the same source, which means that we measure the similarity of two images. The feature values of the two image matrices are used to address this problem. The presence of N feature values is assumed. By comparing the Euclidean distance between the feature vector of the tested finger vein image and a set of feature values of a finger vein image of the p^{th} registrant in the database, the matching result can be obtained through Equation (4).

$$D(F, F_p) = \sqrt{\sum_{i=1}^N (F[i, 1] - F_p[i, 1])^2} \quad (4)$$

In the equation, $D(F, F_p)$ is the Euclidean distance between the feature vector of the finger vein image to be identified and a set of feature vector of a finger vein image of the p^{th} registrant in the database, $F[i, 1]$ is the i^{th} feature values of the feature vector group in the first filtering direction of the finger vein image to be identified, $F_p[i, 1]$ is the i^{th} feature values of the feature vector group in the first filtering direction of the finger vein image of the p^{th} registrant in the database, and N represents the total number of fingerprint features extracted in each filtering direction.

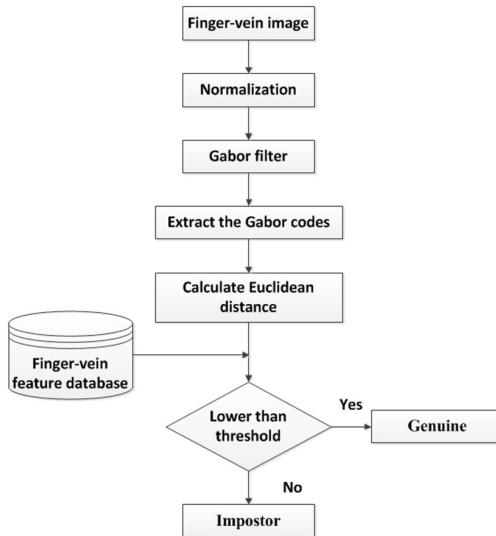
3.4 Implementation of Algorithm

The implementation of the proposed algorithm is shown in Fig. 5. First, we obtain and normalize the captured finger vein image shown in Section 2. Second, the image is filtered by a Gabor filter. Third, feature values are extracted from the image. Euclidean distance is then calculated. Lastly, the Euclidean distance of the finger vein image to be identified and the feature vectors of the image of the registrants in the database are compared. As shown in Equation 4, if $D(f, f_p)$ is lower than the threshold, the match is genuine. In contrast, if $D(f, f_p)$ is higher than the threshold, the match is impostor.

4 Experimental Results

A database of finger vein images is created for experimental evaluation. We collect five finger vein images from each of 50 individuals. A total of 250 finger vein images are obtained. Each finger vein image is then compared with the other samples. Hence, the total number of matches is $250 \times 249 = 62250$, which includes genuine and impostor matching. Genuine matching in this study refers to the comparison of the same fingers of the same person, and impostor matching is the comparison of different fingers.

The parameters of the Gabor filter are selected so that the proposed algorithm achieves the best recognition accuracy on a training dataset. Table 1 lists the recognition and misclassification rates of different parameters, which show that when $f = 1/12$, $\sigma = \sigma_x = \sigma_v = 1.5$ the algorithm exhibits the best performance.

**Fig. 5.** Flowchart of the algorithm**Table 1.** Different filter parameters with corresponding recognition and misclassification rates

σ	f	Recognition rate	Misclassification rate
1.7	1/12	89.2%	0.46%
1.5	1/12	95.6%	0.27%
1.3	1/12	97.0%	11.8%
1.5	1/10	95.3%	2.67%
1.5	1/13	89.0%	0.04%

Genuine and impostor matching are shown in Fig. 6. Figure 6 shows that the central distance between genuine and impostor matching is relatively large. We can effectively distinguish whether the comparison of the finger vein images is genuine or not as long as the selected threshold is reasonable.

The two most important statistical performance indexes in finger vein recognition technology are false acceptance rate (FAR) and false rejection rate (FRR). If abscissa is FAR and longitudinal coordinate is FRR, the ROC curve can be obtained. The point on the ROC curve indicates FRR and FAR at a certain threshold. Figure 7 shows the final ROC curve of finger vein recognition based on Gabor filters. The figure also indicates the point of equal error rate. This point shows that the equal error rate is 0.79% when the threshold is 16. Compared with the equal error rate of 4.8% from previous literature [11], which is based on the improved filtering and correction of the Hausdorff distance in finger vein recognition, the proposed algorithm is significantly more accurate.

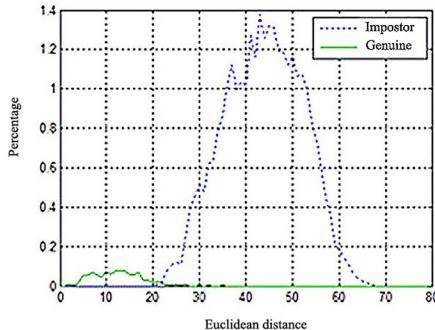


Fig. 6. Distance distribution of genuine and impostor matching

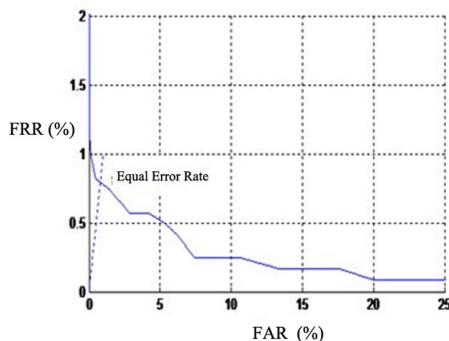


Fig. 7. ROC curve

5 Conclusions

This study proposes a finger vein recognition algorithm based on Gabor filters. The algorithm utilizes a filter in different directions on the normalized image and then extracts feature values from the image after filtering. We then determine the matching level between two images by comparing the Euclidean distance of the feature values. The experiments show that the algorithm has a high recognition rate of 0.79% equal error rate, which is more accurate than the results of existing algorithms.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No. 60977058), Independent Innovation Foundation of Shandong University (IIFSDU) (No.2012JC015 and No. 2012DX001) and the Technology Development Program of Shandong Province (No.2012GGE27073).

References

1. Qichuan, T., Runsheng, Z.: The summary of biometric feature identification. Computer Application Research 26, 4401–4406 (2009)
2. Chengbo, Y., Huafeng, Q.: Biometric feature identification: finger vein identification technology. Beijing Tsinghua University Press (2009)
3. Miura, N., Nagasaka, A., Miyatake, T.: Extraction of finger-vein patterns using maximum curvature points in image profiles. IEICE Transactions on Information and Systems E90D, 1185–1194 (2007)
4. Chengbo, Y., Zhaomin, Z., Hongbing, L., Yanlin, L.: Research on extracting human finger vein pattern characteristics based on residual image. Computer Engineering and Applications 46, 167–169 (2010)
5. Qin, H.F., Qin, L., Yu, C.B.: Region growth-based feature extraction method for finger-vein recognition. Optical Engineering 50, 057208-1–057208-8 (2011)
6. Li, H.B., Yu, C.B., Zhang, D.M., et al.: Study on finger vein image enhancement based on ridgelet transformation. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition 23, 224–230 (2011)
7. Yang, J.F., Yan, M.F.: An improved method for finger-vein image enhancement. In: Proceedings of the 10th IEEE International Conference on Signal Processing (ICSP 2010), Beijing, pp. 1706–1709 (2010)
8. Yang, X., Zhi, L., Zhang, H.X., Zhang, H.: Finger Vein Verification System Based on Sparse Representation. Applied Optics 51, 6252–6258 (2012)
9. Zhang, Y.Y.: Fingerprint image enhancement based on elliptical shape Gabor filter. In: Proceedings of the 6th IEEE International Conference on Intelligent Systems, pp. 344–348 (2012)
10. Wang, Q., Zhang, X.D., Li, M.Q., Dong, X.P., Zhou, Q.H., Yin, Y.: Adaboost and multi-orientation 2D Gabor-based noisy iris recognition. Pattern Recognition Letters 33, 978–983 (2012)
11. Wang, K.J., Ma, H.: Finger vein recognition by improved filtering and correction of Hausdorff distance. Journal of Computer-Aided Design & Computer Graphics 23, 385–391 (2011)

Construction and Simulation Analysis for Stability Speed Parameter of Instantaneous Availability for One-Unit Repairable Systems

Yi Yang^{1,2}, Lichao Wang³, and Rui Kang⁴

¹ Reliability and Systems Engineering School, Beihang University, Beijing 100191

² China Astronaut Research and Training Center, Beijing, China, 100094

zzpcissyy2006@163.com

³ The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing, China
uglyme@yahoo.cn

⁴ Reliability and Systems Engineering School, Beihang University, Beijing 100191
kr@buaa.edu.au

Abstract. According to the stability theory, the mathematical model of the stability parameter of linear systems is established for analysis of instantaneous availability of one-unit repairable systems, and the concept of stability parameter is presented. The conditions of determining the parameter stability are derived, and the measure of parameter stability speed is put forward. On the above basis, the stability speed parameter K is given for one-unit repairable systems in instantaneous availability fluctuation, and the typical fluctuation problems are analyzed by use of simulations. The obtained results confirm the rationality and applicability of the stability speed parameter K.

Keywords: Instantaneous Availability Fluctuation, Fluctuation Stability Speed, Parameter Stability Speed, Stability Speed Parameter.

1 Introduction

How to get a system with high availability is a particularly important task in systems engineering. The reference [1] studied how to improve the reliability and running availability of complex systems, and proposed a method based on pre-diagnosis. A series of simulations were conducted in [1], and the simulation results suggest that system availability fluctuates at the initial stage and then gradually converges to a certain value, as shown in Fig. 1 and Fig. 2.

To improve operational availability, reference [2] performed overhauls at some predetermined frequency, that is, replaced all critical parts at regular intervals. Fig. 1-1 and Fig. 1-2 depict results of the simulation of a system with an overhaul frequency of 137 days and 91.25 days respectively.

As shown in Fig. 1, the instantaneous availability of the two platforms with different overhaul frequencies first undulates and then stabilizes to a steady-state value. Similar phenomena can be found in [3]. Actually, the phenomena of the instantaneous availability undulation abound in practical engineering. For example, the F-15, the aircraft of U.S. military, had an undulated instantaneous availability at the stage of

initial operation. In 1980s, the aircraft's instantaneous availability was about 37.5% in a combat exercise, while it was 9% in a peacetime training. However, the availability of F-15 reached a high level of 93.7% in Gulf War in 1990s. Obviously, the availability of the aircraft F-15 showed great fluctuations over the period of 10 years.

Actually, the instantaneous availability of many complex weapons systems exhibits a certain range of fluctuations at the beginning of use, which cannot be characterized by the steady-state availability or interval availability^{[4]-[6]}, and the most commonly used indicators in effectiveness evaluation are often unable to describe this kind of undulation. The references [7]-[8] investigated the instantaneous availability fluctuations, and a set of characteristic parameters were proposed to describe the instantaneous availability fluctuations, such as adaptive time, index function, availability amplitude and so on.

These characteristic parameters play a very important role in describing the instantaneous availability fluctuations, but they are dependent on the calibration level, which is subjectively determined by the people. Therefore, it is necessary to introduce some new parameters to provide strong theoretical support for better characterization of instantaneous availability fluctuations. Based on the stability theory, this paper builds up a parameter model of stability speed. Then a series of simulations are carried out to verify the rationality and applicability of the new parameter.

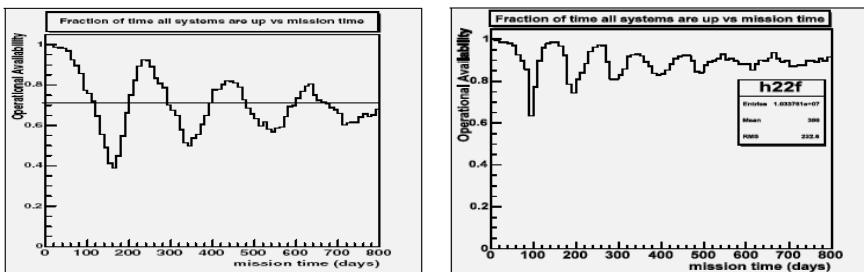


Fig. 1. Availability of the platform with overhaul frequency of 137 days and 91.25 days

2 Parameter Stability Speed

2.1 Definition of Parameter Stability

The *maximum* Assume that the system dynamic equation is given by

$$x(k+1) = f(x(k), u(k)) \quad (1)$$

the output equation is given by

$$y(k) = g(x(k)) \quad (2)$$

the performance parameter is

$$r(k) = h(x(k)) \in R^l \quad (3)$$

and the initial state of the system is

$$x(0) = x_0 \in \Omega \subset R^n \quad (4)$$

where

$x(k) \in R^n$ denotes the system state,

$u(k) \in R^l$ denotes the system input,

$y(k) \in R^m$ denotes the system output,

$f(\cdot)$ denotes the system dynamic function,

$g(\cdot)$ denotes the system output function,

$h(\cdot)$ denotes the parameter - state relation function

The system input/output block diagram is shown in Fig. 3.

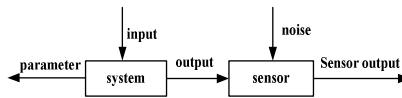


Fig. 2. Input and output block diagram

We first start with the simplest case. Consider the linear system described by

$$x(k+1) = Ax(k) + b \quad (5)$$

$$y(k) = Cx(k) + d \quad (6)$$

$$r(k) = Px(k) + e \quad (7)$$

$$x(0) = x_0 \in \Omega \quad (8)$$

Where $A \in M_n$ denotes the system matrix, $C \in M_{m,n}$ denotes the output matrix, $b \in R^n$, $d \in R^m$, $e \in R^l$ denote some constant vectors.

Definition 1: For the system (5)-(8), given the parameter stability domain requirement Θ , if there is a constant $K_0 \geq 0$ such that

$$r(k) \in \Theta, \quad \forall k \geq K_0$$

then the parameter $r(k)$ is said to be stable, or the parameter is said to be bounded stable. When $r(k)$ is stable, it is also called the bounded stability parameter. Furthermore, for a given norm $\|\cdot\|$, if $\exists r^* \in R^l$, for $\forall \epsilon > 0$, $\exists K_1 \geq 0$ such that

$$\|r(k) - r^*\| < \epsilon, \quad \forall k \geq K_1$$

then $r(k)$ is called asymptotically stable with the equilibrium point being r^* , that is,

$$r(k) \rightarrow r^*, \quad k \rightarrow \infty$$

Obviously, if the parameter $r(k)$ is asymptotically stable, we can get the conclusion that it is also bounded stable.

Definition 2: Define $T = \max\{k \mid r(k) \notin \Theta\}$ as the parameter fluctuation period. The system parameter becomes stable when the time k satisfies $k > T$. Generally, the system stability domain is expressed as

$$\Theta = \left\{ x \in R^d \mid \|x - r^*\| \leq Q \right\}$$

where $Q > 0$ is a given value.

2.2 Decision Condition of Parameter Stability

Before studying the various conditions of parameters stability, we give a few lemmas.

Lemma 1 [9]: Assume that $A \in M_n$, and $\varepsilon > 0$ is a given number. Then there exists a matrix norm $\|\cdot\|$ satisfying

$$\rho(A) \leq \|A\| \leq \rho(A) + \varepsilon$$

Where $\rho(A) = \max\{|\lambda| : \lambda \text{ is the characteristic value of } A\}$ is the spectral radius of matrix A .

Lemma 2 [10]: Assume that $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two vector norms on the finite-dimensional real or complex vector space V . Then there exist two constants $C_m, C_M > 0$, such that

$$C_m \|x\|_\alpha \leq \|x\|_\beta \leq C_M \|x\|_\alpha \quad \forall x \in V$$

Theorem1: For the system (5)-(8) and a given constant $Q > 0$, if the spectral radius of the system matrix A is less than 1, then the parameter $r(k)$ is asymptotically stable.

Proof: Since the spectral radius of the system matrix A is less than 1, the matrix $I - A$ is invertible. Moreover, the system (5)-(8) is stable, and the equilibrium point is $x^* = (I - A)^{-1} b$. Let

$$r^* = P(I - A)^{-1} b + e,$$

Because $\rho(A) < 1$, according to Lemma 1, there exists an induced norm $\|\cdot\|_\alpha$ satisfying

$$\rho(A) \leq \|A\|_\alpha \leq \rho(A) + \frac{1 - \rho(A)}{2} = \frac{1 + \rho(A)}{2} < 1$$

Then, we can get that

$\|r(k) - r^*\|_\alpha = \|Px(k) - Px^*\|_\alpha \leq \|P\|_\alpha \|A\|_\alpha \|x(k-1) - x^*\| \leq \dots \leq \|P\|_\alpha \|A\|_\alpha \|x(0) - x^*\|_\alpha$ According to Lemma 2, for any norm $\|\cdot\|$, there exists a constant $C_M > 0$ such that

$$\|x\| \leq C_M \|x\|_\alpha, \quad \forall x \in V$$

Then

$\|r(k) - r^*\| \leq C_M \|r(k) - r^*\|_\alpha \leq C_M \|P\|_\alpha \|A\|_\alpha^k \|x(0) - x^*\|_\alpha$ By Definition 1, we can conclude that the parameter $r(k)$ is asymptotically stable. The proof is complete.

Furthermore, assume that matrix A has n linearly independent eigenvectors $\alpha_1, \alpha_2, \dots, \alpha_n$, and the corresponding eigenvalues are $\lambda_1, \lambda_2, \dots, \lambda_n$ ($|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$, and $\lambda_i \neq 1$). So there exist a series of constant s_1, s_2, \dots, s_n , such that

$$x_0 - x^* = s_1 \alpha_1 + s_2 \alpha_2 + \dots + s_n \alpha_n \quad (9)$$

where

$$x^* = (I - A)^{-1} b$$

is the system equilibrium point. Let

$$r^* = P(I - A)^{-1} b + e$$

We have that

$$\begin{aligned} r(k) - r^* &= P[x(k) - x^*] = PA[x(k-1) - x^*] = \dots = PA^k [x(0) - x^*] \\ &= PA^k (s_1 \alpha_1 + s_2 \alpha_2 + \dots + s_n \alpha_n) \\ &= P(s_1 A^k \alpha_1 + s_2 A^k \alpha_2 + \dots + s_n A^k \alpha_n) \\ &= s_1 \lambda_1^k P \alpha_1 + s_2 \lambda_2^k P \alpha_2 + \dots + s_n \lambda_n^k P \alpha_n \end{aligned} \quad (10)$$

Therefore, in order to have

$$r(k) \rightarrow r^*, \quad k \rightarrow \infty$$

the following condition must be satisfied:

$$s_1 \lambda_1^k P \alpha_1 + s_2 \lambda_2^k P \alpha_2 + \dots + s_n \lambda_n^k P \alpha_n \rightarrow 0, \quad k \rightarrow \infty$$

Then, we have the following result.

Theorem 2: For the system (5)-(8), assume that $\alpha_1, \alpha_2, \dots, \alpha_n$ are the linearly independent eigenvectors of the system matrix A and the corresponding eigenvalues are $\lambda_1, \lambda_2, \dots, \lambda_n$, ($|\lambda_1| \geq \dots \geq |\lambda_j| \geq 1 > |\lambda_{j+1}| \geq \dots \geq |\lambda_n|$, $(0 \leq j \leq n)$. If it is satisfied that

$$s_1 P \alpha_1 = 0, s_2 P \alpha_2 = 0, \dots, s_j P \alpha_j = 0$$

then the parameter $r(k)$ is asymptotically stable.

Theorem 2 shows that the parameter stability is related to the system stability, the initial state of the system, and the parameter-state relation.

2.3 Measurement of Parameter Stability Speed

Some criterions of parameter stability are given in the preceding subsection. For the stable parameters, two questions naturally arise: How to depict the stability speed? How to accelerate the stability speed? In order to address these two questions, we need to find a measure to characterize the parameter stability speed.

It follows from Theorem 2 that

$$r(k+1) - r^* = s_{j+1} \lambda_{j+1}^{k+1} P\alpha_{j+1} + s_{j+2} \lambda_{j+2}^{k+1} P\alpha_{j+2} + \dots + s_n \lambda_n^{k+1} P\alpha_n$$

$$r(k) - r^* = s_{j+1} \lambda_{j+1}^k P\alpha_{j+1} + s_{j+2} \lambda_{j+2}^k P\alpha_{j+2} + \dots + s_n \lambda_n^k P\alpha_n$$

Then

$$\begin{aligned} \|r(k+1) - r^*\| &= \left\| s_{j+1} \lambda_{j+1}^{k+1} P\alpha_{j+1} + s_{j+2} \lambda_{j+2}^{k+1} P\alpha_{j+2} + \dots + s_n \lambda_n^{k+1} P\alpha_n \right\| \\ \|r(k) - r^*\| &= \left\| s_{j+1} \lambda_{j+1}^k P\alpha_{j+1} + s_{j+2} \lambda_{j+2}^k P\alpha_{j+2} + \dots + s_n \lambda_n^k P\alpha_n \right\| \\ |\lambda_{j+1}| &\left\| s_{j+1} P\alpha_{j+1} + s_{j+2} \left(\frac{\lambda_{j+2}}{\lambda_{j+1}} \right)^{k+1} P\alpha_{j+2} + \dots + s_n \left(\frac{\lambda_n}{\lambda_{j+1}} \right)^{k+1} P\alpha_n \right\| \\ &\left\| s_{j+1} P\alpha_{j+1} + s_{j+2} \left(\frac{\lambda_{j+2}}{\lambda_{j+1}} \right)^k P\alpha_{j+2} + \dots + s_n \left(\frac{\lambda_n}{\lambda_{j+1}} \right)^k P\alpha_n \right\| \end{aligned} \quad (11)$$

Obviously, $\frac{\|r(k+1) - r^*\|}{\|r(k) - r^*\|}$ is related to $\frac{\lambda_{j+2}}{\lambda_{j+1}}, \frac{\lambda_{j+3}}{\lambda_{j+1}}, \dots, \frac{\lambda_n}{\lambda_{j+1}}$, especially to $|\lambda_{j+1}|$.

Therefore, from the perspective of convenient engineering applications, we choose $|\lambda_{j+1}|$ to characterize the parameters stability speed.

3 Analysis of Parameter Stability Speed for Instantaneous Availability

Assume that the system consists of one-unit [11], and the failure time X follows the discrete distribution

$$p_k = P\{X = k\} \quad k = 0, 1, 2, \dots$$

The unit will be repaired immediately after its failure, and the system repaired can work just like a new one. Let the repair time Y obey the discrete distribution

$$q_k = P\{Y = k\} \quad k = 0, 1, 2, \dots$$

Define the system state as

$$\begin{cases} Z(k) = 0 & \text{the system is normal at } k; \\ Z(k) = 1 & \text{the system is being repaired at } k; \end{cases} \quad k = 0, 1, 2, \dots$$

Without loss of generality, it is assumed that $P\{Z(0) = 0\} = 1$.

The corresponding system failure rate and repair rate are respectively

$$\lambda(k) \in (0, 1), k = 0, 1, 2, \dots, n_1 - 1, \quad \lambda(n_1) = 1$$

$$\omega(k) \in (0, 1), k = 0, 1, 2, \dots, n_2 - 1, \quad \omega(n_2) = 1$$

Then we can obtain the system state transition equation

$$P(k+1) = BP(k) \quad k = 0, 1, 2, \dots \quad (12)$$

Its instantaneous availability is $A(k) = \delta_{n_1+1, n_2+1} P(k)$

$$(13)$$

where

$$B = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 & \infty(0) & \infty(1) & \dots & \infty(n_2-1) & 1 \\ 1-\lambda(0) & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1-\lambda(1) & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \dots & \dots \\ 0 & 0 & \dots & 1-\lambda(n_1-1) & 0 & 0 & 0 & \dots & 0 & 0 \\ \lambda(0) & \lambda(1) & \dots & \lambda(n_1-1) & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1-\infty(0) & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1-\infty(1) & \dots & 0 & 0 \\ \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1-\infty(n_2-1) & 0 \end{bmatrix}$$

$$P(k) = (P_0(k,0), P_0(k,1), \dots, P_0(k,n_1), P_1(k,0), P_1(k,1), \dots, P_0(k,n_1))^\top \in [0,1]^{n_1+n_2+2}$$

$$\delta_{m,n} = \left(\underbrace{\overbrace{1, \dots, 1}^m, \overbrace{0, \dots, 0}^n} \right)$$

Let $PV(k) = (P_0(k,0), P_0(k,1), \dots, P_0(k,n_1), P_1(k,0), P_1(k,1), \dots, P_0(k,n_1-1))^\top \in [0,1]^{n_1+n_2+1}$ We can get that

$$PV(k+1) = B_1 PV(k) + B_2 \quad (14)$$

Where

$$B_1 = \begin{bmatrix} -1 & -1 & \dots & -1 & -1 & \infty(0)-1 & \infty(1)-1 & \dots & \infty(n_2-2)-1 & \infty(n_2-1)-1 \\ 1-\lambda(0) & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1-\lambda(1) & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \dots & \dots \\ 0 & 0 & \dots & 1-\lambda(n_1-1) & 0 & 0 & 0 & \dots & 0 & 0 \\ \lambda(0) & \lambda(1) & \dots & \lambda(n_1-1) & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1-\infty(0) & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1-\infty(1) & \dots & 0 & 0 \\ \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1-\infty(n_2-2) & 0 \end{bmatrix}$$

$$B_2 = (1, 0, \dots, 0)^\top = \delta_{1,n_1+n_2}^T \in R^{n_1+n_2+1}$$

Let

$$PV^* = (I - B_1)^{-1} B_2$$

By the analysis made in the previous section, $K = |\lambda_i(B_1)|$ is chosen to describe the parameter stability speed. Obviously, it is easy to get that

$$\lim_{k \rightarrow \infty} \frac{\|PV(k+1) - PV^*\|}{\|PV(k) - PV^*\|} \leq K$$

4 Simulation Study

Assume that the system failure rate and repair rate are respectively

$$\lambda(k) = 1 - q_1^{(k+1)\beta_1 - k\beta_1}, \quad k = 0, 1, 2, \dots, n_1 - 1, \quad \lambda(n_1) = 1$$

$$\infty(k) = 1 - q_2^{(k+1)\beta_2 - k\beta_2}, \quad k = 0, 1, 2, \dots, n_2 - 1, \quad \infty(n_2) = 1$$

Let $n_1 = n_2 = 595$. According to the different values of other parameters, we do the following groupings.

The first group: $q_1 = 0.998, \beta_1 = 3; q_2 = 0.98, \beta_2 = 3$

Then we can get that $|\lambda_i(B_1)| = 0.9046$

The second group: $q_1 = 0.998, \beta_1 = 2; q_2 = 0.98, \beta_2 = 2$

Then we can get that $|\lambda_i(B_1)| = 0.9142$

The third group: $q_1 = 0.998, \beta_1 = 2; q_2 = 0.98, \beta_2 = 2$

Then we can get that $|\lambda_1(B_1)| = 0.9978$

The fourth group: $q_1 = 0.998$, $\beta_1 = 1$; $q_2 = 0.998$, $\beta_2 = 1$

Then we can get that $|\lambda_1(B_1)| = 0.9980$

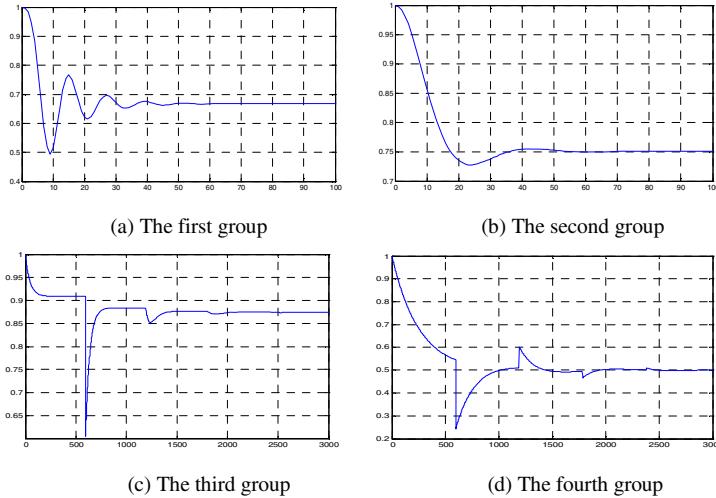


Fig. 3. Availability of the system

As shown in Fig. 3, it is seen that the availability fluctuations period extends as the norm of the largest eigenvalue of the system matrix increases, which is in good agreement with the theoretical analysis made in Section 2.3. Therefore, it is reasonable and appropriate to use the norm of the largest eigenvalue to describe the availability fluctuations.

5 Conclusion

This paper has introduced the concept of the parameter stability, and proposed the criterion to determine the parameter stability. On this basis, the availability speed has been adopted to describe the availability fluctuations. Computer simulations have been carried out. The simulation results have indicated that the availability stability speed can depict the stability characteristics of the fluctuations of the instantaneous availability. This has provided a new way to study the fluctuation mechanism of instantaneous availability. The future work includes the study of the influences of the reliability-time parameters, maintainability-time parameters and supportability-time parameters onto the availability stability speed, which will constitute the basis of the mechanism analysis of instantaneous availability fluctuations.

This instruction file for Word users (there is a separate instruction file for LaTeX users) may be used as a template. Kindly send the final and checked Word and PDF files of your paper to the Contact Volume Editor. This is usually one of the organizers

of the conference. You should make sure that the Word and the PDF files are identical and correct and that only one version of your paper is sent. It is not possible to update files at a later stage. Please note that we do not need the printed paper.

We would like to draw your attention to the fact that it is not possible to modify a paper in any way, once it has been published. This applies to both the printed book and the online version of the publication. Every detail, including the order of the names of the authors, should be checked before the paper is sent to the Volume Editors.

References

- [1] Operational Requirements Document (ORD) for the Future Combat Systems, Prepared by UAMBL, For Knox, Kentucky (2003)
- [2] Macheret, Y., Koehn, P., Sparrow, D.: Improving reliability and operational availability of military systems. In: 2005 IEEE Aerospace Conference, Montana, USA, pp. 3489–3957 (2005)
- [3] Pham-Gia, T., Turkkan, N.: System availability in a Gamma alternating renewal process. Naval Research Logistics 46, 822–844 (1999)
- [4] Carrasco, J.A.: Solving large interval availability models using a model transformation approach. Computers & Operations Research 31, 807–861 (2004)
- [5] Claasen, S.J., Joubert, J.W., Yadavalli, V.S.S.: Interval estimation of the availability of a two-unit standby system with non-instantaneous switch-over and 'dead time'. Pakistan Journal of Statistics 20, 115–122 (2004)
- [6] Kirmani, E., Hood, C.S.: A new approach to analysis of interval availability. In: The Third International Conference on Availability, Reliability and Security, pp. 479–483 (2008)
- [7] Wang, L.C.: The Analysis and Design of the Matchinglization for System Availability. Nanjing University of Science & Technology (2009)
- [8] Wang, L.C., Yang, Y., Yu, Y.L., Zou, Y.: Analysis of matchable problems based on system availability. Journal of Systems Engineering 24, 253–256 (2009)
- [9] Lei, J.G., Tang, P., Tian, R.: Matrix Theory and Application. Mechanical Industry Press, Beijing (2005)
- [10] Horn, R.A., Johnson, C.R., Yang, Q.Y.: Matrix Analysis. Mechanical Industry Press, Beijing (2005)
- [11] Tang, Y.H., Liu, X.Y.: A New One Unit Repairable System. Systems Engineering-Theory and Practice 23, 106–111 (2003)

Compressed Sensing Ensemble Classifier for Human Detection

Baochang Zhang¹, Juan Liu¹, Yongsheng Gao², and Jianzhuang Liu^{3,4}

¹ Science and Technology on Aircraft Control Laboratory,
School of Automation Science and Electrical Engineering, BeiHang University,
Beijing, 100191, China

² School of Engineering, Griffith University, Australia

³ Shenzhen Key Lab for Computer Vision and Pattern Recognition,

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

⁴ Department of Information Engineering, The Chinese University of Hong Kong, China
bczhang@buaa.edu.cn

Abstract. This paper proposes a novel Compressed Sensing Ensemble Classifier (CSEC) for human detection. The proposed CSEC employs the compressed sensing technique to get a more sparse model with a more reasonable selection of base classifiers. The major contributions of this paper are: 1) a novel principled framework for ensemble classifier design based on compressed sensing; 2) a new concept of considering both the simplicity of ensemble classifier and irrelevance of base classifiers towards optimal classifier design; and 3) a quadratic function for CSEC optimization which includes a new optimizable positive semi-definite relevance matrix to simultaneously select appropriate base classifiers with minimized relevance. Experimental results on INRIA and SDL databases show that the performance of CSEC is better than two most popular classifiers SVM and AdaBoost, as well as a most recent method CLML.

Keywords: compressed sensing, ensemble, classifier.

1 Introduction

Following the large amount of research work on face detection, human detection has naturally become the next popular research topic due to its wide range of applications in video surveillance, content-based image retrieval and video annotation. However, the task of human detection is very challenging because of the existence of significant variations of pose, clothing, illumination, and cluttered background.

Many interesting human detection approaches have been proposed in the literature. For example, Mohan et al. [1] adopt silhouette information to represent humans, and use SVM for final classification. Mu et al. [2] also propose a pedestrian detection system implemented by SVM. Following the Haar-like feature based AdaBoost cascade [3] for face detection, Tuzel et al. [4] first transform features into a tangent space of Riemannian manifolds, and then use the LogitBoost cascade for

classification. In [5], SVM is employed with Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) [6] features, and good human detection accuracy is achieved. Zhu et al. [7] use linear-SVM as weak classifiers and build an AdaBoost cascade mechanism for human detection. Munder and Gavrila [8] conduct an experimental study on pedestrian classification. They conclude that SVM performs the best, and AdaBoost achieves comparable performance at a much lower computational cost. Recently, Xu et al. [9] propose a human detection method based on Cascaded L1-norm Minimization Learning (CLML) in the boosting framework, and obtain very good human detection performance.

This paper proposes a novel Compressed Sensing Ensemble Classifier (CSEC) for human detection. The CSEC approach is based on a principled framework and thus different from AdaBoost. CSEC is also different from SVM due to its use of Compressed Sensing (CS) instead of margin maximization to get a more sparse model[12]. CSEC is theoretically based on Occam's razor [13,14], one of the fundamental principles in the theory of machine learning. This principle states that the best model prefers the smallest complexity, i.e., in order to achieve low prediction error, the classifier should be as simple as possible. As the VC dimension is commonly used as the measure of complexity in machine learning, therefore, among all the capable classifiers, the optimal one prefers the minimum VC-dimension. In this sense, CSEC actually provides a new effective and direct way of minimizing the VC-dimension of classifier, and thus towards optimal classifier design.

The rest of this paper is organized as follows: Section 2 introduces the two principles used in the CSEC framework. Section 3 presents the detailed procedure of how to implement the two-layer cascade of CSEC classifier. Section 4 gives our experimental evaluation of the proposed approach on the application of human detection. Section 5 concludes this paper and gives some discussions and future work.

2 Principles of CSEC

For clarity, in the following discussions we refer to each of the classifiers that are combined in the vote as a base classifier and refer to the final voted classifier as the ensemble classifier. The proposed Compressed Sensing Ensemble Classifier (CSEC) is formulated as:

$$\begin{aligned} \min_{w, \varepsilon_t} \quad & \|w\|_1 + C_1 w^T R w + C_2 \sum_t \varepsilon_t \\ \text{s.t.} \quad & \text{constraints of } w \end{aligned} \tag{1}$$

where w is the vector formed by all the weights assigned to each of the base classifiers. R is a specially defined relevance matrix to represent the relationship/relevance between base classifiers. Detailed information of R is explained in the next section. ε_t is an introduced positive slack variable to find a soft margin. C_1 and C_2 are two positive parameters to balance three components in (1), i.e., the sparseness, the irrelevance and the training error.

3 Implementation of CSEC

The proposed CSEC approach is implemented in two layers of cascade. Only a positive result from the first layer of cascade triggers the second layer. In the first layer, CSEC is learnt with linear base classifiers. Using linear base classifiers is computationally efficient with acceptable detection rates. In the second layer, CSEC is extended to feature selection to form a non-linear local kernel classifier. Using a non-linear classifier is computationally expensive with high detection rates. By using the cascaded CSEC, the system obtains a trade-off between speed and performance.

3.1 The Linear CSEC Classifier

3.1.1 Linear Base Classifier Generation

Given a labeled training set $\{x_t, y_t\}$, $x_t \in \Re^d$, $y_t \in \{-1, +1\}$ and $t = 1, \dots, N$, the base classifiers are learned by L1-norm minimization:

$$\begin{aligned} \min_{p_k, \varepsilon_t} \quad & \|p_k\|_1 + C_1 \sum_{t=1}^N \varepsilon_t \\ \text{s.t.} \quad & \begin{cases} y_t \cdot h_k(x_t) \geq 1 - \varepsilon_t \\ \varepsilon_t \geq 0, \quad t = 1, 2, \dots, N \end{cases} \end{aligned} \quad (2)$$

where N is the number of training samples, p_k is the normal vector of the k -th linear base classifier $h_k(x_t) = p_k^T x_t - a_k$, y_t is the class label of x_t , ε_t is the introduced positive slack variable, and C_1 is a predefined parameter to balance the training error and L1-norm minimization. The larger C_1 is, the smaller the training error is forced to be (i.e., fewer misclassified samples). To minimize the misclassification rate in the learning process C_1 is assigned a larger value (50.0 in our experiments).

3.1.2 Relevance Matrix Definition

Each base classifier has different classification ability, and each pair of different classifiers have different compensation ability against each other. Therefore, we define a specially designed relevance matrix R to represent the relationship/relevance between base classifiers. The elements $R_{i,j}$ of the relevance matrix are defined differently for $i \neq j$ and $i = j$ as follows:

$$R_{i,j} \triangleq \begin{cases} \eta \exp(-\|v_i - v_j\|_2^2 / 2\sigma^2), & i \neq j \\ 1 - v_i^T v_i / N, & i = j \end{cases} \quad (3)$$

where $\sigma > 0$, $\eta = 1/\sqrt{2\pi}\sigma$, v_i and v_j are the binary classification results on the training dataset from the i -th and j -th base classifiers, respectively. Here we use the classifier performance (represented by the binary classification results) to calculate the compensability of different classifiers. If the classification is correct, the binary

classification result is 1; otherwise, it is 0. For different base classifiers, $R_{i,j}$ ($i \neq j$) is calculated by the Gaussian function. A larger value of $R_{i,j}$ indicates more relevance (or less compensability) between the i -th and j -th base classifiers. In other words, the two classifiers have very similar classification results on the training dataset. Therefore, only one of the two base classifiers is expected to be included in the ensemble classifier. On the contrary, a smaller value of $R_{i,j}$ indicates less relevance (or more compensability) between two classifiers. In this case, both two base classifiers are expected to be selected.

3.1.3 CSEC Optimization

Based on the above definitions, the elements of the relevance matrix $R_{i,j}$ represent either the compensability between two different base classifiers (when $i \neq j$) or the classification ability of one base classifier (when $i = j$). An optimal set of base classifiers is supposed to have smallest $R_{i,j}$. To minimize the relevance matrix, we assign a weight variable to each of the base classifiers. The weight w_i assigned to the i -th base classifier directly controls whether (or to what extent) the classifier should be included in the ensemble classifier. If $R_{i,j}$ is large (poor compensability between the i -th and j -th base classifiers when $i \neq j$ or poor classification ability of the i -th base classifier when $i = j$), its associated weight should be minimized. The weights can be optimized through:

$$\min_{w_i} \sum_i \sum_j w_i w_j R_{i,j} \quad (4)$$

which includes the combined minimization of the relevance between different base classifiers and the error rate of the base classifiers.

Following the CSEC learning principles, we add the L1-norm minimization into the optimization function. Then the selection of base classifiers is optimized through:

$$\begin{aligned} \min_{w, \varepsilon_t} & \|w\|_1 + C_1 \sum_i \sum_j w_i w_j R_{i,j} + C_2 \sum_{t=1}^N \varepsilon_t \\ & = \|w\|_1 + C_1 w^T R w + C_2 \sum_{t=1}^N \varepsilon_t \\ \text{s.t. } & \begin{cases} y_t \cdot (\sum_k w_k f_k - b) \geq 1 - \varepsilon_t \\ \varepsilon_t \geq 0, \quad t = 1, 2, \dots, N \end{cases} \end{aligned} \quad (5)$$

where w is the vector formed by all the weights w_k assigned to each base classifier. f_k is the output of the k -th base classifier on the training set and represented by the base classifier's classification ability ($f_k \in [0,1]$). b is a threshold. ε_t is an introduced positive slack variable to find a soft margin. C_1 and C_2 are two positive parameters to balance three components in (5).

In practice, however, the relevance matrix R is not guaranteed to be positive definite or positive semi-definite, and therefore cannot necessarily form an easily

solvable convex optimization problem. To address this issue, we reformulate the minimization part of (5) into:

$$\begin{aligned} \min_{w, \epsilon_t} \quad & \|w\|_1 + C_1(w^T R w + w^T A I w) + C_2 \sum_{t=1}^N \epsilon_t \\ = & \|w\|_1 + C_1 w^T (R + A I) w + C_2 \sum_{t=1}^N \epsilon_t \end{aligned}, \quad (6)$$

where I is the identity matrix with the same size of R , and A is a scalar value. When $A \geq \eta$ (see (3)), it can be proved that $R + A I$ is positive semi-definite (see the next section). By minimizing (6) instead of (5), the optimization problem can be easily solved using many convex optimization algorithms (see [20] for example). We can prove $R + A I$ being Positive Semi-Definite as shown in our extended version in mpl.buaa.edu.cn.

3.2 The Nonlinear CSEC Classifier

For complex object detection, nonlinear classifiers generally demonstrate better performance than linear ones. At the second layer of CSEC cascade, we design a nonlinear classifier to further refine some false positive results from the first layer. The nonlinear classifier is constructed by extending CSEC to the kernel based feature selection and learnt on the misclassified training samples from the first layer. Given a training image divided into multiple overlapping or non-overlapping parts, let z_i represent the feature vectors extracted from the i -th image part. We assign a weight variable s_i to each of the local kernel features to control whether (or to what extent) the features of the i -th part should be included in the nonlinear classifier. Following the relevance matrix definition in CSEC, the weights can be optimized through minimizing the relevance/similarity between the features of different image parts:

$$\min_{s_i} \quad \sum_i \sum_j s_i s_j \kappa_{i,j} \quad (7)$$

where $\kappa_{i,j}$ is the local kernel function, used as a trick to calculating the inner products in the high-dimensional space projected by some implicit mapping function $\Phi(\bullet)$:

$$\kappa_{i,j} = \kappa(z_i, z_j) = \Phi(z_i)^T \Phi(z_j). \quad (8)$$

Here, we use the Gaussian kernel:

$$\kappa_{i,j} = \kappa(z_i, z_j) = \eta \exp(-\|z_i - z_j\|_2^2 / 2\sigma^2). \quad (9)$$

It can be easily seen that the local kernel function $\kappa_{i,j}$ gets a large value when z_i and z_j is highly correlated. In this case, only one of the two image parts is expected to be included in the nonlinear classifier for efficiency. Following the CSEC learning and combining (7) with the L1-norm minimization, we get:

$$\begin{aligned}
\min_{s, \epsilon_t} \quad & \|s\|_1 + C_1 \sum_i \sum_j s_i s_j \kappa_{i,j} + C_2 \sum_{t=1}^N \epsilon_t \\
& = \|s\|_1 + C_1 s^T \kappa s + C_2 \sum_{t=1}^N \epsilon_t \\
\text{s.t.} \quad & \begin{cases} y_t \cdot (\sum_i s_i \sum_j \kappa(z_i, z_j)) \geq 1 - \epsilon_t \\ \epsilon_t \geq 0, \quad t = 1, 2, \dots, N \end{cases}
\end{aligned} \tag{10}$$

where s is the vector formed by all the weights s_i assigned to each of the local kernel features.

The nonlinear CSEC classifier is similar to kernel trick based nonlinear SVM classifier. However, due to the complexity of human body, SVM is very computationally expensive for human detection. Therefore, CSEC in fact constructs a light version of local kernel nonlinear classifier from image parts via compressed sensing as well as redundancy removal between image parts to increase the computational efficiency of the system for human detection.

4 Experiments

The proposed CSEC method is evaluated on the challenging INRIA database [6] and SDL database [21] for the application of human detection. These test sets cover diverse body poses and complex backgrounds, while most humans are in standing position. Although the positive training samples in our experiments are mostly humans of frontal view, the CSEC approach demonstrates noticeable capability of handling multi-posture and occlusion cases in our experiments. For CSEC training, we extract 5000 positive training samples from the SDL database, and 5000 negative large training samples from the INRIA database.

4.1 Feature Extraction

In our experiments, the CSEC learning is based on the commonly used HOG features [6] extracted from each of the 64×128 human detection windows. The HOG descriptor can be considered as a local contour representation of objects. It captures statistical histogram features from dense grids of gradient or edge information around human body. In the original design of HOG, a 64×128 detection window is divided into multiple overlapping blocks of size 16×16 , and each block consists of $2 \times 2 = 4$ cells with size of 8×8 . Gradient orientations of pixels in a cell are projected into discrete 9 orientation feature bins. Thus each block is represented by a 36-dimensional vector. Details of the feature extraction procedure can be found in [6].

4.2 Performance Comparison

The proposed CSEC approach is compared with three state-of-the-art benchmark methods, HOG+SVM [6], HOG+AdaBoost [7] and CLML [9], all using the HOG descriptor to represent humans. The comparative detection accuracies against false

positives per window (FPPW) on the INRIA database and the SDL database are shown in Figs. 1, respectively. It can be seen from the figures that CSEC obtains a better performance than the benchmark methods on both databases. On the INRIA database, CSEC achieves a much better performance than HOG+SVM and HOG+AdaBoost, and a slightly better performance than CLML. At the FPPW rate of 10^{-5} , CSEC performs about 10% higher than HOG+SVM, and about 4% higher than HOG+AdaBoost. On the SDL database, CSEC achieves a much better performance than CLML (about 5%). These comparative results confirm that CSEC is a powerful classifier for human detection.

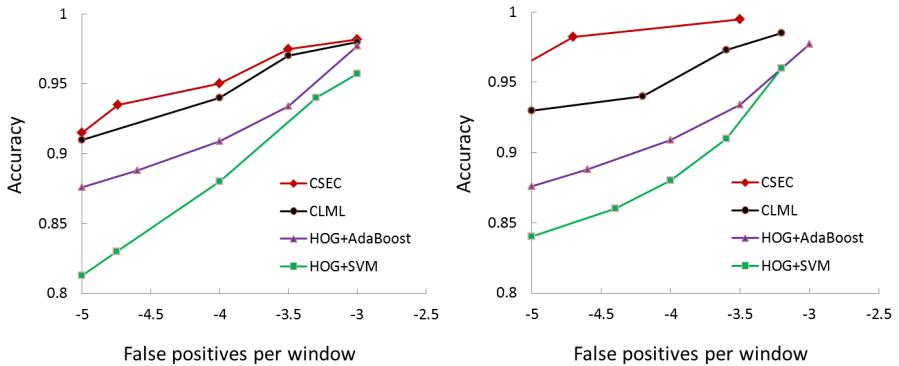


Fig. 1. The comparative results on the INRIA database (left) and SDL database (right)

5 Conclusions

We propose a new direction for the computer vision research community by demonstrating that the compressed sensing technique can be used for optimal ensemble classifier design as well as feature selection. The proposed CSEC classifier is formulated as an easily solvable convex quadratic optimization problem, which achieved a better performance in accuracy than CLML, SVM and AdaBoost for human detection. More importantly, CSEC performs significantly efficient in computational speed than other benchmark methods, demonstrating the compressed sensing concept is working. This further reflects Occam's razor that the best model prefers the smallest complexity. Although it is perhaps unsafe to claim that CSEC is the optimal ensemble classifier, it is without doubt going towards that way.

Acknowledgement. This work was supported in part by the Natural Science Foundation of China, under Contracts 60903065, 61039003, 61070148 and 61272052, in part by the Fundamental Research Funds for the Central Universities, and in part by Guangdong Innovative Research Team Program (No.201001D0104648280), and the Program for New Century Excellent Talents in University of Ministry of Education of China. Jianzhuang Liu is the corresponding author. Baochang Zhang thank Yao Cao for helping with the experiment and the algorithm implementation. Thanks for Ran Xu from Graduate school of Chinese Academy Sciences providing us the test platform in the experiments.

References

1. Mohan, A., Papageorgiou, C., Poggio, T.: Example-Based Object Detection in Images by Components. *IEEE Trans. PAMI* 23(4), 349–360 (2001)
2. Mu, Y., Yan, S., Liu, Y., Huang, T., Zhou, B.: Discriminative Local Binary Patterns for Human Detection in Personal Album. In: Proc. CVPR (2008)
3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR (2001)
4. Tuzel, F.P., Meer, P.: Human Detection via Classification on Riemannian Manifolds. In: Proc. CVPR, pp. 1–8 (2007)
5. Wang, X., Han, T.X., Yan, S.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: Proc. ICCV, Kyoto (2009)
6. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proc. CVPR, vol. 1, pp. 886–893 (2005)
7. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: Proc. CVPR, vol. 2, pp. 1491–1498 (2006)
8. Munder, S., Gavrila, D.: An Experimental Study on Pedestrian Classification. *IEEE Trans. PAMI* 28(11), 1863–1868 (2006)
9. Xu, R., Zhang, B., Ye, Q., Jiao, J.: Cascaded L1-norm Minimization Learning (CLML) classifier for human detection. In: Proc. CVPR (2010)
10. Platt, J.C.: Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In: NIPS, pp. 557–563 (1998)
11. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2(2), 121–167 (1998)
12. Donoho, D.: For most large underdetermined systems of linear equations the minimal l1-norm near solution approximates the sparsest solution. *Comm. on Pure and Applied Math.* 59(6), 797–829 (2006)
13. Thorburn, W.M.: Occam's razor. *Mind* 24, 287–288 (1915)
14. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust Face Recognition via Sparse Representation. *IEEE Tran. PAMI* 31, 210–227 (2009)
15. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
16. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990)
17. Bartlett, P., Freund, Y., Lee, W.S., Schapire, R.E.: Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *Ann. Statist.* 26(5), 1651–1686 (1998)
18. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: Many could be better than all. *Artificial Intelligence* 137(1-2), 239–263 (2002)
19. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
20. Grant, M.: Disciplined convex programming. PhD thesis, Stanford (2004)
21. <http://coe.gucas.ac.cn/SDL-HomePage/resource.asp>

A Prediction Reference Structure Based Hierarchical Perceptual Encryption Algorithm for H.264 Bitstream

Haojie Shen, Li Zhuo, and Yirui Li

Signal & Information Processing Lab, Beijing University of Technology,
Beijing 100124, China

13810761814@139.com, zhuoli@bjut.edu.cn,
liyirui@mails.bjut.edu.cn

Abstract. In this paper, correspondence between the degree of video motion and the motion reference ratio (MRR) of macroblock (MB) has been studied firstly. An efficient hierarchical perceptual encryption algorithm for H.264 bitstream is proposed based on MRR. First, at frame layer, frames to be encrypted are selected dynamically according to the degree of motion. Then, at MB layer, MBs to be encrypted are selected based on their MRRs. And finally, at bitstream layer, the most significant bits for reconstructed video quality are encrypted on the basis of the bit-sensitivity of H.264 bitstream. Furthermore, MB layer has a quality control factor to control the number of encrypted MBs, thus a fine-grained visual quality control can be achieved. Experimental results demonstrate that the proposed algorithm can provide sufficient security, and achieve wide-range quality control with low computational complexity. Furthermore, it has no impact on the compression ratio and can maintain the format-compliance.

Keywords: Perceptual video encryption, visual quality control, motion reference ratio, bit-sensitivity, RC4.

1 Introduction

With the development of network multimedia applications, the security of video information has become increasingly prominent. In recent years, a variety of video encryption algorithms have been put forward, which can be divided into complete encryption and selective encryption [1]. The latter is the mainstream video encryption method, and can achieve a tradeoff between security and computational complexity.

In some important applications such as video conferencing, encryption algorithm should ensure that no visual information can be visible. They are called “fully confidential” video encryption algorithms [2-4]. On the contrary, in many applications for entertainment, such as video-on-demand (VoD), the video signal should be encrypted only to a certain extent, and some video contents are still perceptible. Such low-quality version of the original video can serve as a preview for attracting potential customers to subscribe the high-quality version. This leads to the so-called perceptual video encryption [5-7].

For perceptual video encryption, it is desirable that the visual quality degradation can be continuously controlled by a control factor P , which generally represents a

percentage corresponding to the encryption strength. In addition, $P \in [0,1]$ and a larger P value means stronger encryption strength. Generally, an efficient perceptual encryption algorithm should also support format-compliance, remain the compression ratio unchanged, have a low computation complexity and be easy to implement, while may not need to resist some complex attacks because the cost of applying such attacks is usually higher than paying for the services.

Perceptual encryption algorithms for various video coding standards such as MPEG-4 and H.264 have been proposed recently. Li et al. [5] proposed a perceptual video encryption algorithm named PVEA for MPEG-4 that encrypts fix-length codewords with three different control parameters. However, this method is not available for the bitstream encoded mostly by variable length coding (VLC) such as H.264 bitstream. Wang et al. [6] proposed to encrypt the coded_block_pattern (CBP), signs of trailing ones and levels of nonzero coefficients for H.264. However, such algorithm results in lower compression efficiency. Au-Yeung et al. [7] achieved the perceptual encryption at the transform stage during the encoding process, i.e. applied one of multiple transforms which were developed to be used as alternates to DCT to individual blocks according to a randomly generated key. However, to some extent, compression efficiency will be affected by using other transforms instead of DCT.

As can be seen, the existing perceptual video encryption algorithms are usually incorporated into the encoding process and reduce the compression efficiency, as well as need to modify the structure of the standard encoder. Furthermore, these algorithms are based on the coding framework and the syntax structure of video bitstream. The most important information for reconstructed video quality is selected to be encrypted, such as DCT coefficients and motion vector differences (MVDs). Therefore, the encryption locations are often fixed and video content is ignored. However, video content is of great variety. The location-fixed encryption scheme will lead to the poor pertinence of the perceptual encryption.

Therefore, in this paper, an efficient hierarchical perceptual encryption algorithm in H.264 compressed domain is proposed. First, frames are dynamically selected to be encrypted based on the degree of motion. Then, MBs are selected to be encrypted based on their MRRs. And finally, the most significant bits for reconstructed video quality are selected to be encrypted based on the bit-sensitivity of H.264 bitstream. MB layer has a quality control factor to control the number of encrypted MBs, which provides users with fine-grained and wide-range visual quality control, and thereby can meet the different quality requirements of applications.

2 The Correspondence between MRR and the Degree of Motion

H.264 standard uses motion estimation/compensation to remove temporal redundancy. As shown in Fig. 1, Pixel a in Frame n is used as reference by Pixel b and c in Frame $n+1$, which will be further used by Pixel d , e and f in Frame $n+2$. This reference structure will continue within the current group of pictures (GOP) until the next I-frame (Intra-frame). Obviously, if Pixel a is changed by encryption, Pixel $b\sim z$ will all be changed as well. Consequently, using this reference dependency can significantly improve the efficiency of encryption.

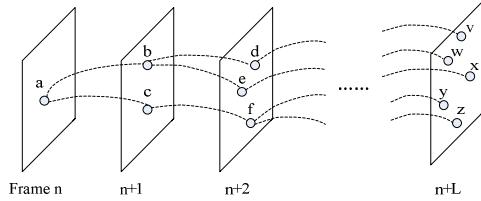


Fig. 1. The reference map of pixels among video frames

MRR is proposed to represent the degree of inter-frame dependency. As in [8], the MRR $\gamma(n, m)$ for the m^{th} MB in Frame n is defined as follows:

$$\gamma(n, m) = \frac{1}{K} \sum_{(i, j) \in MB(n, m)} M_{i,j}(n) \quad (1)$$

where $M_{i,j}(n)$ is the motion reference map (MRM) of Frame n , which denotes the total number of pixels in subsequent frames which use Pixel $p(i,j)$ in Frame n as motion prediction reference directly or indirectly. $MB(n, m)$ is the m^{th} MB in Frame n , and K is the number of pixels in one MB. For H.264 encoded video, $K=16 \times 16$.

Obviously, the MRR has a strong correlation with the degree of motion of video sequence. In order to verify the relationship between MRR and the degree of motion, eight standard video sequences of QCIF format and GOP=15 were tested, including “Football”, “Bus”, “Foreman”, “Mobile”, “Highway”, “Hall”, “Mother” and “Akiyo”.

The average values and the standard deviations of MRRs for each frame of the test video sequences are shown in Fig. 2. As can be seen, when the degree of motion of video sequence is higher, the average value of MRRs is smaller while the standard deviation of MRRs is larger. Therefore, the following conclusions can be drawn: 1) A small average value and a large standard deviation of MRRs correspond to a video with higher degree of motion; 2) MBs with smaller MRR in one frame usually correspond to regions whose motion is relatively drastic. Then, an efficient hierarchical perceptual video encryption algorithm is proposed by using MRR to select frames and MBs for encryption.

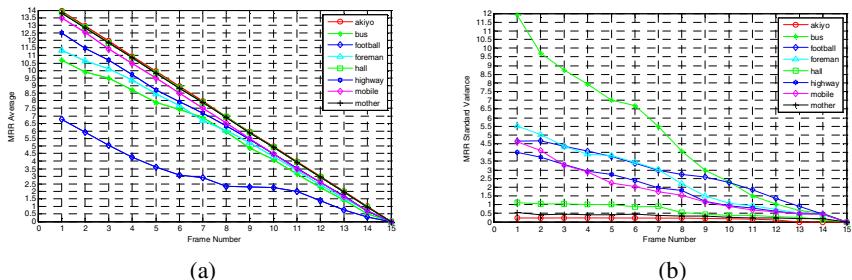


Fig. 2. The average values and the standard deviations of MRRs for each frame of the test video sequences, (a) average values, (b) standard deviations

3 The Proposed Perceptual Video Encryption Algorithm

The general framework of the proposed perceptual encryption algorithm can be outlined as follows: Firstly, the data to be encrypted are determined according to the frame-layer, MB-layer and bitstream-layer selection strategy and control factor P_M . Secondly, the selected bits for encryption are extracted and concatenated to form a new substream. Thirdly, the substream is encrypted with RC4 stream cipher. And finally, the encrypted bits are put back into their original positions to form the encrypted H.264 stream. The proposed algorithm directly encrypts the compressed bitstream so that it has no impact on the compression ratio.

3.1 Selection of Frames to Be Encrypted

Fig. 3 shows the objective quality of the first GOP of different video sequences with the number of encrypted frames changed, where structural similarity (SSIM) [9] is used as quality assessment metric. It can be seen that, with the reduction of the degree of motion, encrypting subsequent P-frames will have less effect on the visual quality degradation. Therefore, in order to improve the efficiency of encryption, P-frames of each GOP are selected to be encrypted dynamically based on the degree of motion. To be precise, MRR is used to determine the number of P-frames to be encrypted (N_{P_enc}) based on the following formula:

$$N_{P_enc} = \begin{cases} (N_{GOP} - 1)\sqrt{\sigma_I/E_I}, & \sigma_I/E_I \leq 1 \\ N_{GOP} - 1, & \sigma_I/E_I > 1 \end{cases} \quad (2)$$

where N_{GOP} is the number of frames in the GOP. σ_I and E_I are the standard deviation and the average value of MRRs of I-frame respectively.

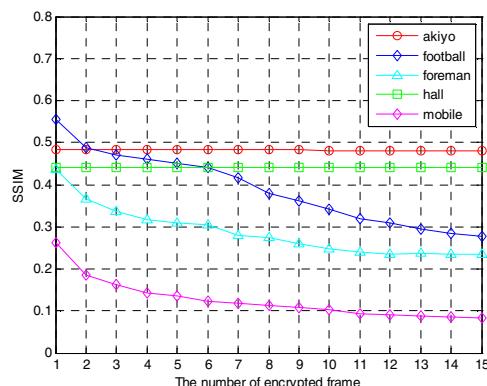


Fig. 3. The objective quality of “Football”, “Foreman”, “Mobile”, “Hall” and “Akiyo” with the number of encrypted frames changed

3.2 Selection of MBs to Be Encrypted

In this paper, each MB is classified into two categories based on its MRR, as shown in the following formula:

$$e(n,m) = \begin{cases} 1, & \gamma(n,m) < T(n) \\ 0, & \gamma(n,m) \geq T(n) \end{cases} \quad (3)$$

where $e(n,m)=1$ denotes that the m^{th} MB in Frame n will be classified to be motion-dramatic, and $e(n,m)=0$ refers the opposite. $T(n)$ is the average value of MRRs for Frame n .

Thus, based on the classification result, for the videos with gentle motion, only MBs of I-frames will be encrypted due to the small displacement of MBs between adjacent frames. The associated content in the subsequent frames can still be imperceptible by the strong reference dependency among frames. For the videos with drastic motion, considering that the displacement of MBs between adjacent frames is large, the MBs in both of I and P frames should be encrypted.

To sum up, the proposed method selects the following MBs for encryption: All MBs of I-frames and MBs of P-frames whose MRR is smaller than the average value.

3.3 Selection of Bits to Be Encrypted

Bit-sensitivity [4] is adopted to select bits for encryption, which represents the degree of importance of each bit in the compressed bitstream for reconstructed video quality.

Through a great number of experiments, we have found that for H.264 bitstream, the corrupted coding parameters such as IntraPredMode, MVDSign, MVDLevel and LowCoeffSign of the I-frame will cause the most significant decrease of PSNR values. In other words, these parameters are the most important bits for video reconstruction.

In order to improve the encryption efficiency, the intra-prediction mode codewords and the sign bits of the low frequency DCT coefficients are selected to be encrypted for intra-coded MBs, while the info_suffix of MVD codewords are encrypted for inter-coded MBs. Furthermore, in order to keep the encrypted bitstream format-compliant, only the syntax element rem_intra4x4_pred_mode and the last one bit of the intra16x16 codewords are encrypted for the encryption of intra-prediction mode.

3.4 Quality Control

Obviously, the MB whose MRR is larger is more important for the video quality. Thus, MRR is used as importance weight of each MB for visual quality, and the quality control can be achieved as follows:

$$s(n) = P_M \cdot \sum_{MB(n,m) \in S} \gamma(n,m) \quad (4)$$

where $s(n)$ is the sum of MRRs of encrypted MBs after quality control for Frame n . $P_M \in [0,1]$ is the control factor, S is the set of selected MBs according to Section 3.2.

$\dot{\gamma}(n,m) = \gamma(n,m) + 1$ is the modified MRR, i.e. $MB(n,m)$ itself should be added. Furthermore, each frame is divided into two parts according to the Formula (3) and all MBs of I-frames are selected for encryption. Therefore, for the I-frames, the quality of two parts is controlled separately.

4 Experiments and Performance Analyses

In order to demonstrate the effectiveness of the proposed perceptual video encryption algorithm, various video sequences of CIF and QCIF format were tested in our experiments. All the video sequences are encoded by JM86 (H.264 codec software) baseline profile with GOP=15. Each encoded sequence contains 150 frames.

4.1 Visual Quality

The reconstructed frames of “Mobile”, “Forman” and “Football” using the proposed encryption algorithm are shown in Fig. 4, where P_M is 0.1, 0.3, 0.6 and 0.9 respectively. As can be seen, the visual quality of all video sequences declines (from slightly blurred to imperceptible) as P_M increases. A high perceptual security has been achieved when $P_M=0.9$. Consequently, the proposed perceptual encryption algorithm can provide wide-range quality control.

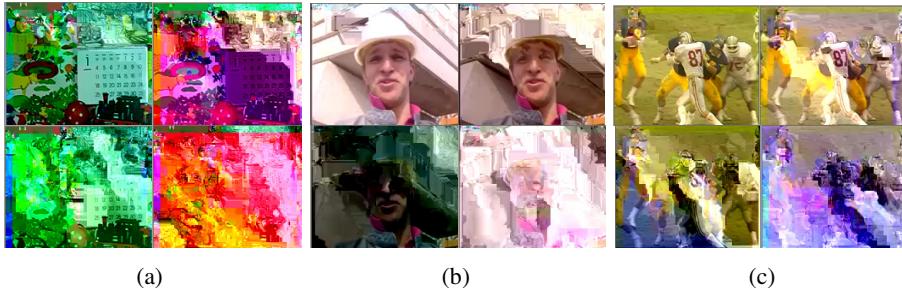


Fig. 4. (a)(b)(c) are the reconstructed frames of “Mobile”, “Forman” and “Football” respectively (Top Left: $P_M=0.1$, Top Right: $P_M=0.3$, Bottom Left: $P_M=0.6$, Bottom Right: $P_M=0.9$)

4.2 Security

The replacement attack is aimed to recover the encrypted information and make it more visually perceptible by replacing the encrypted data with other data. In our experiments, the encrypted intra-prediction modes are all replaced by the most_probable_mode (the minimum of the prediction modes of neighboring blocks) and the encrypted MVD signs are all replaced by those of adjacent blocks. The recovered frames of “Football”, “Foreman” and “Mobile” (encrypted with $P_M=0.6$) after replacement attack are shown in Fig. 5. It can be seen that the recovered frames become more chaotic, which proves the security against replacement attack.



Fig. 5. The recovered frames of “Football”, “Foreman” and “Mobile” after replacement attack by replacing the encrypted intra prediction modes and MVD signs

4.3 Computational Complexity

The data-selection result of each layer with the largest encryption strength is shown in Table 1, including the ratio of encrypted frames to total frames (EFR), encrypted MBs to total MBs (EMBR), and encrypted bits to the entire bitstream (EBR). As can be seen, the encrypted data volume of the proposed algorithm is around only 50% of that of Ref. [4], which means that the efficiency of encryption is significantly improved by multi-layer selection while security can also be fully guaranteed.

Table 2 gives the time-efficiency of the encryption with the largest encryption strength, which is measured by the ratio between encryption time and compression time (ECTR). It can be seen that the ECTR of the proposed algorithm is not more than 1% except “Football”, which shows that compared with compression, the time consumption of the encryption process is negligible. In conclusion, the proposed algorithm is of lower computational complexity and will be well suitable for real-time video applications.

Table 1. The data-selection result with the largest encryption strength ($P_M = 1$)

Size	Video	EFR	EMBR	EBR	
				The proposed algorithm	Ref. [4]
QCIF	Akiyo	21%	8.80%	7.62%	14.02%
	Silent	46%	16.86%	6.81%	14.16%
	Foreman	74%	38.77%	6.41%	12.76%
CIF	Mobile	62%	41.58%	2.84%	14.46%
	Tempete	57%	28.38%	3.35%	7.03%
	Football	96%	64.20%	9.09%	18.09%

Table 2. The time-efficiency of the encryption with the largest encryption strength (ECTR)

Size	Videos	ECTR	Size	Video	ECTR
QCIF	Akiyo	0.24%	CIF	Mobile	0.77%
	Silent	0.38%		Tempete	0.59%
	Foreman	0.45%		Football	1.44%

5 Conclusions

An efficient hierarchical perceptual encryption algorithm for H.264 bitstream is proposed in this paper. First, at frame layer, P-frames of each GOP to be encrypted are selected dynamically based on the degree of motion. Then, at MB layer, MBs are selected to be encrypted based on their MRRs, including all MBs of I-frames and the MBs of P-frames whose MRR is smaller than the average value. And finally, at bitstream layer, the intra-prediction mode codewords, the sign bits of the low frequency DCT coefficients and the info_suffix of the MVD codewords are selected for encryption. Meanwhile, a quality control factor is used to control the number of encrypted MBs. Experimental results demonstrate that the proposed perceptual encryption algorithm can provide sufficient security, and achieves wide-range quality control for all video sequences with low computational complexity.

Acknowledgments. The work in this paper is supported by the Program for New Century Excellent Talents in University (No.NCET-11-0892), the Specialized Research Fund for the Doctoral Program of Higher Education (No.20121103110017), the National Natural Science Foundation of China (No.61003289, No.61100212), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD201304036).

References

1. Liu, F., Koenig, H.: A Survey of Video Encryption Algorithms. *J. Computers and Security* 29(1), 3–15 (2010)
2. Wu, C.P., Kuo, C.C.J.: Design of Integrated Multimedia Compression and Encryption Systems. *IEEE Transactions on Multimedia* 7(5), 828–839 (2005)
3. Lian, S., Liu, Z., Ren, Z., Wang, H.: Secure Advanced Video Coding Based on Selective Encryption Algorithms. *IEEE Transactions on Consumer Electronics* 52(2), 621–629 (2006)
4. Zhuo, L., Mao, N.S., Zhang, J., Li, X.G.: Bit-Sensitivity Based Video Encryption Scheme in Compressed Domain. *IJACT: International Journal of Advancements in Computing Technology* 4(8), 155–164 (2012)
5. Li, S., Chen, G., Cheung, A., Bhargava, B., Lo, K.: On the Design of Perceptual MPEG-Video Encryption Algorithms. *IEEE Transactions on Circuit and Systems for Video Technology* 17(2), 214–223 (2007)
6. Wang, F., Wang, W., Ma, J., Xiao, C., Wang, K.: Perceptual Video Encryption Scheme for Mobile Application Based on H. 264. *The Journal of China Universities of Posts and Telecommunications* 15(1), 73–78 (2008)
7. Au-Yeung, S.K., Zhu, S., Zeng, B.: Partial Video Encryption Based on Alternating Transforms. *IEEE Signal Processing Letters* 16(10), 893–896 (2009)
8. Zhang, Y., Qin, S., He, Z.: Fine-Granularity Transmission Distortion Modeling for Video Packet Scheduling over Mesh Networks. *IEEE Transactions on Multimedia* 12(1), 1–12 (2010)
9. Au-Yeung, S.K., Zhu, S., Zeng, B.: Quality Assessment for A Perceptual Video Encryption System. In: *IEEE International Conference on Wireless Communications, Networking and Information Security*, pp. 102–106 (2010)

A New Image Denoising and Enhancement Method Combining the Nonsubsampled Contourlet Transform and Improved Total Variation

Ying Li, Yu Jia, and Yanning Zhang

School of Computer Science, Northwestern Polytechnical University,
Shaanxi, Xi'an 710129, China
lybyp@nwpu.edu.cn, Jiay423@163.com

Abstract. Transform-based denoising methods are very popular in recent years. However, they often suffer from unwanted artifacts like pesudo-Gibbs phenomena. In this paper, we propose a new hybrid image denoising by combining the nonsubsumpled contourlet transform (NSCT) with improved total variation. First, an improved stark function which integrates noise reduction with feature enhancement is developed to nonlinearly shrink and stretch the NSCT coefficients. Then an improved Total variation is introduced to reduce the pseudo-Gibbs artifacts of the enhanced image which are caused by the elimination of small NSCT coefficients. Numerical experiments show that this approach improves the image quality by enhancing the shape of edges and important detailed features while suppressing noise in comparison to many well known methods.

Keywords: image denoising, image enhancement, nonsubsumpled contourlet transform, total variation.

1 Introduction

Image denoising and enhancement is an import pre-processing task for further processing of image like segmentation, feature extraction, texture analysis etc. Denoising and enhancement refers to suppressing the noise while retaining or even extrude the edges and other important detailed structures as much as possible. The tools to resolve this problem come from very different fields like computational harmonic analysis (CHA) and partial differential equations (PDEs).

Contourlet transform was constructed by Donoho et al. in 2002[1][2], then the nonsubsumpled contourlet transform was proposed by Cunha et al. in 2006[3]. It is a “real” two-dimensional image representation, which can capture the intrinsic geometrical structure of the image. Denoising using the nonsubsumpled contourlet transform has got better results than wavelet and some other methods, but at the same time, it suffers from unwanted artifacts when we use a threshold to suppress the noise of the image, which effects the quality of the image. So how to remove the artifacts out from the denoised image becomes a problem worthy of studying.

On the other hand, the partial differential equations (PDEs) have been extensively applied over the last two decades in signal and image processing [4-7]. Total variation image restoration model, as a kind of PDEs, was first proposed by Rudin et al. [8,9], and is an available method to solve ill-posed problems in image processing. In this paper, we propose a method which uses total variation to reduce the oscillations in NSCT thresholding approximations because many people have proved that total variation reduce the pseudo-Gibbs artifacts well [10].

In this paper, a combined approach for noisy image enhancement by using NSCT and total variation is introduced. Firstly an improved Starck function is applied to NSCT coefficients for image enhancement and noise reduction, and afterwards a total variation process is performed to the enhanced and denoised result in order to reduce the pseudo-Gibbs artifacts, which caused by the elimination of thresholded NSCT coefficients.

2 Brief Review of NSCT

The contourlet transform is a real two-dimensional transform using a cascade of Laplacian pyramid and a directional filter bank, it can capture intrinsic geometric structure information of images and achieve better expression than discrete wavelet transform, especially for edges and contours. Moreover, it is easily adjustable for detecting fine details in any orientation along curvatures, which results in more potential for effective analysis of images. The contourlet transform can efficiently capture the smooth contours in the natural texture images.

However, the contourlet transform is lack of shift-invariance because of the downsampling and upsampling, so Cunha et al. proposed nonsubsampled contourlet transformation in 2006. It consists of two filter banks, nonsubsampled pyramid filter bank(NSPFB) and nonsubsampled directional filter bank(NSDFB), the NSPFB provides nonsubsampled multi-scale decomposition and captures the point discontinuities. The NSDFB provides nonsubsampled directional decomposition and links point discontinuities into linear structures.

2.1 The Nonsubsampled Pyramid Filter Bank

The multiscale property of the NSCT is obtained from a shift-invariant filtering structure that achieves a subband decomposition similar to that of laplacian pyramid. The nonsubsample filter bank decomposition generates a lowpass version and a highpass version of the original image. The process is then iterated on the coarse version. In the end, a lowpass and several bandpass versions are obtained and a nonsubsample pyramid structure is realized.

2.2 The Nonsubsampled Directional Filter Bank

A shift-invariant directional expansion is obtained by a NSDFB. The NSDFB is constructed by eliminating the downsamplers and upsamplers in the DFB. This is done

by switching off the downsamplers upsamplers in each two-channel filter bank in the DFB tree structure and upsampling the filters accordingly. This results in a tree composed of two-channel nonsubsampled filter banks.

3 Improved Total Variation

Among the various noise suppressing methods, the total variation (TV) has been quite successful due to its properties that the intensity change is related to the scale of the features in the image. TV models are based on functional analysis and differential geometry, which can be formulated as follows:

$$\arg \min \left\{ \lambda \int_{\Omega} |\nabla u| d\Omega + \frac{1}{2} \int_{\Omega} (u - u_0)^2 d\Omega \right\} \quad (2)$$

where u is the result image, and u_0 denotes the original image. Total variation minimization model of energy is composed by the regularization term and the fidelity term. The regularization term makes the image become smooth while the fidelity term plays an important role in preserving edges and details. In our paper, In order to preserve even enhance the wanted oriented textures, we use the improved TV model, that is:

$$\frac{\partial u}{\partial t} = (u - Su_0) - \lambda \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) \quad (3)$$

the nonlinear operator S should preserve and enhance the important features of the image. Here we propose to use a transform based on wave atoms shrinkage method. Because Oscillatory functions or oriented textures have a significantly sparser expansion in wave atoms than in other fixed standard representations like Gabor filters, wavelets, and curvelets. The fidelity term then can be written as:

$$(u - Su_0) = u - (WA)^{-1} \theta WA(u_0) \quad (4)$$

where WA denotes the transform of wave atoms.

4 Combining NSCT and Improved TV for Image Enhancement

There has been a rapidly increasing interest in designing hybrid methods using both multi-resolution analysis method shrinkage and TV denoising methods. For the similar goal, The aim of the hybrid method proposed in this paper is to resolve the contradiction between noise suppression, pseudo-Gibbs removing and texture preserving, which cannot be resolved by the TV-based method or wavelet method independently.

Firstly we apply the forward Nonsubsampled contourlet transform (NSCT) to the noisy image I_0 , we obtain NSCT coefficient $c_{j,l,k}(I_0) = NSCT(I_0)$. Then, an

improved Starck function is used to modify the NSCT coefficients to get the enhanced image. The improved Starck function is:

$$y_c(x, \sigma) = \begin{cases} 0 & \text{if } x < c\sigma \\ \frac{x-c\sigma}{c\sigma} \left(\frac{m}{2c\sigma} \right)^p + \frac{2c\sigma-x}{c\sigma} & \text{if } c\sigma \leq x < 2c\sigma \\ \left(\frac{m}{x} \right)^p & \text{if } 2c\sigma \leq x < m \\ \left(\frac{m}{x} \right)^s & \text{if } x \geq m \end{cases} \quad (5)$$

where p determines the degree of nonlinearity and s introduces dynamic range compression. Using a nonzero s will enhance the faintest edges and soften the strongest edges at the same time. σ is the estimated noise standard deviation, c is a normalization parameter, and a c value larger than 2 guarantees that the noise will not be amplified. The m parameter is the value under which coefficients are amplified, $m = k\sigma$, k is a const.

By using the improved gain function y_c for adjusting $c_{j,l,k}(I_0)$, an enhanced image can be expressed as

$$I_{con} = \sum_{j>1} y_c(|c_{j,l,k}(I_0)|, c\sigma) \cdot c_{j,l,k}(I_0) \cdot \varphi_{j,l,k} + \sum_{j=1} c_{j,l,k}(I_0) \cdot \varphi_{j,l,k} \quad (6)$$

Thus, an enhanced result can be obtained from nonlinearly shrink and stretch the NSCT coefficients alternatively. The enhanced image however contains pseudo-Gibbs artifacts due to the elimination of small NSCT coefficients. A total variation process is then introduced in order to reduce the pseudo-Gibbs phenomenon.

Now we deal with the enhanced image I_{con} as a cartoon-like image I_{con}^* plus a residual image I_{tv} containing textures and artifacts, i.e., $I_{con} = I_{con}^* + I_{tv}$. In order to calculate I_{tv} , here a so-called inverse thresholding function is used,

$$r(x, T) = \begin{cases} 0, & x \geq T \\ 1, & x < T \end{cases} \quad (7)$$

where $T = c\sigma$. The NSCT coefficients of I_{tv} can be expressed as

$$c_{j,l,k}(I_{tv}) = r(|c_{j,l,k}(I_{con})|, T) \cdot c_{j,l,k}(I_{con}). \quad (8)$$

Then apply the inverse NSCT transform (INSCT) to $c_{j,l,k}(I_{tv})$, the diffusion image is obtained by

$$I_{dif} = INSCT(c_{j,l,k}(I_{tv})) \cdot \quad (9)$$

The above process can be described with a operator $P_v(\cdot)$,

$$I_{tv} = P_v(I_{con}) = INSCT(r(|c_{j,l,k}(I_{con})|, T) \cdot c_{j,l,k}(I_{con})) \cdot \quad (10)$$

Consider the diffusion process

$$\frac{\partial I}{\partial t} = (P_v(I) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I)}{|\nabla P_v(I)|} \right) \quad (11)$$

and the smoothed image is defined as

$$Tv(I) = I^{(t+1)} = I^{(t)} + (P_v(I^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I^{(t)})}{|\nabla P_v(I^{(t)})|} \right) \quad (12)$$

We can get

$$Tv(I_{con}) = I_{con}^{(t+1)} = I_{con}^{(t)} + (P_v(I_{con}^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I_{con}^{(t)})}{|\nabla P_v(I_{con}^{(t)})|} \right) \quad (13)$$

with $I_{con}^0 = I_{con}$, and

$$Tv(I_{tv}) = I_{tv}^{(t+1)} = I_{tv}^{(t)} + (P_v(I_{tv}^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I_{tv}^{(t)})}{|\nabla P_v(I_{tv}^{(t)})|} \right) \quad (14)$$

with $I_{tv}^0 = I_{tv}$.

For $t=0$, we have $I_{con}^0 = I_{con} = I_{con}^* + I_{tv} = I_{con}^* + I_{tv}^0$. Now, assuming $I_{con}^{(t)} = I_{con}^* + I_{tv}^{(t)}$, followed by $P_v(I_{con}^{(t)}) = P_v(I_{con}^* + I_{tv}^{(t)}) = P_v(I_{tv}^{(t)})$, we can get

$$\begin{aligned} Tv(I_{tv}) &= I_{tv}^{(t+1)} = I_{tv}^{(t)} + (P_v(I_{tv}^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I_{tv}^{(t)})}{|\nabla P_v(I_{tv}^{(t)})|} \right) \\ &= I_{con}^{(t)} - I_{con}^* + (P_v(I_{tv}^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I_{tv}^{(t)})}{|\nabla P_v(I_{tv}^{(t)})|} \right) \\ &= I_{con}^{(t)} - I_{con}^* + (P_v(I_{con}^{(t)}) - SP_v(I_0)) - \lambda \nabla \cdot \left(\frac{\nabla P_v(I_{con}^{(t)})}{|\nabla P_v(I_{con}^{(t)})|} \right) \\ &= I_{con}^{(t+1)} - I_{con}^* = Tv(I_{con}) - I_{con}^* \end{aligned} \quad (15)$$

According to the inductive method, we get the conclusion that $I_{con}^t = I_{con}^* + I_{tv}^t$.

Thus the final enhanced and denoised image can be obtained by

$$I_{enh} = T\nu(I_{con}) = T\nu(I_{tv} + I_{con}^*) = T\nu(I_{tv}) + I_{con}^* \quad (16)$$

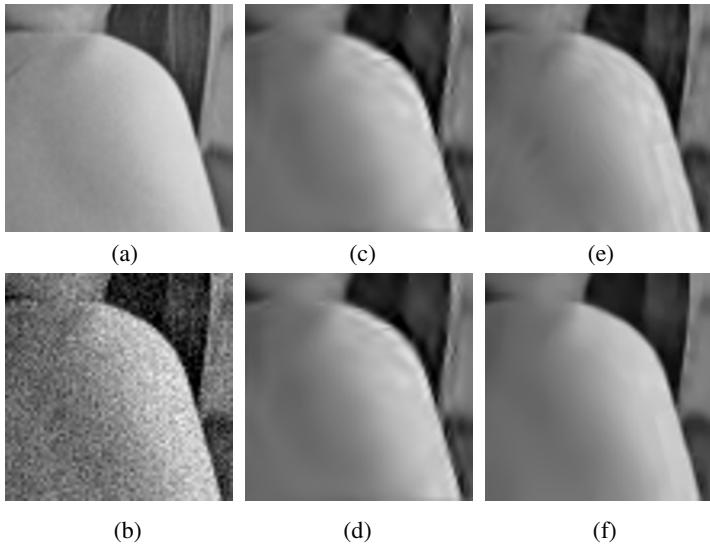
5 Experimental Results

The proposed method is implemented and tested on various test images and its performance is compared with that of many other existing techniques including the wavelet-based enhancement method with/without total variation and NSCT-based enhancement method. To further evaluate the performance of the different enhancement methods objectively, we add Gaussian noise with different standard deviation to the original image, then obtain the enhanced images using four methods and the corresponding edge images using a Canny edge detector. The performance of enhancement algorithm can be quantified by using two quality indices: a noisy suppression quality index (ρ) and an edge preservation index (FOM). If the measured image is close to the reference image, the values of ρ and FOM should be close to 1.

The example is the image of Lena which contains both textures and smooth regions. we show performances of the proposed method in comparison to other three methods. We can observe that the edges are preserved well while the artifacts are suppressed. It is presented that the performance of the enhancement method is improved well through embedding nonlinear diffusion.



Fig. 1. Denoising and enhancement of lena. (a) Original image. (b) Noisy image corrupted by Gaussian noise with $\sigma=15$. (c) Enhanced image by wavelet method. (d) Enhanced image by wavelet_tv method. (e) Enhanced image by NSCT method. (f) ours.

**Fig. 2.** (a)-(f) Close-up of Lena derived from Fig. 1(a)-(f)**Table 1.** *FOM* and ρ Comparison for Different Methods

	Test Images	Noisy	Wav	NSCT	Wav_TV	proposed
Lena	<i>FOM</i>	0.6995	0.8441	0.9125	0.8637	0.9601
	ρ	0.5728	0.9835	0.9876	0.9838	0.9878

6 Conclusion

In this paper, a nonsubsampled contourlet transform-based enhancement method combined with an improved total variation process for noisy image enhancement is proposed. First, an improved Starck function which integrates noise reduction with feature enhancement is developed to nonlinearly shrink and stretch the NSCT coefficients. Then the improved total variation is introduced to reduce the pseudo-Gibbs artifacts of the enhanced image which are caused by the elimination of small NSCT coefficients. Experimental results of the proposed method show better performance of noise reduction, features enhancement and the pseudo-Gibbs artifacts elimination than wavelet-based, wavelet-based diffusion, and NSCT-based enhancement methods. The improved total variation approach has been reflected a significant adventure of the suppression of the pseudo-Gibbs phenomenon.

Acknowledgements. This work was supported by the Research Fund for the Doctoral Program of Higher Education (No. 20126102110041), the Aeronautical Science Foundation of China (No. 2011ZD53049, No. 20125153025).

References

1. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans. Image Process.* 14(12), 2091–2106 (2005)
2. Do, M.N., Vetterli, M.: Contourlets: A directional multiresolution image representation. In: Proc. IEEE Int. Conf. Image Processing (2002)
3. Cunha, A.L., Zhou, J., Do, M.N.: ‘The nonsubsampled contourlet transform: theory, design and applications. *IEEE Trans. Image Process.* 15(10), 3089–3101 (2006)
4. Alvarez, L., Guichard, F., Lions, P.-L., Morel, J.-M.: Axioms and fundamental equations of image processing. *Arch. Rational Mech. Anal.* 123, 199–257 (1993)
5. Catte, F., Lions, P., Morel, M., Coll, T.: Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* 29(3), 182–193 (1992)
6. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259–268 (1992)
7. Rudin, L., Osher, S.: Total variation based image restoration with free local constraints. In: Proc. 1st IEEE Int. Conf. Image Processing, vol. 1, pp. 31–35 (1994)
8. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259–268 (1992)
9. Rudin, L., Osher, S.: Total variation based image restoration with free local constraints. In: Proc. 1st IEEE Int. Conf. Image Processing, vol. 1, pp. 31–35 (1994)
10. Gilboa, G., Sochen, N., Zeevi, Y.: Variational denoising of partly textured images by spatially varying constraints. *IEEE Trans. Image Process.* 15(8), 2281–2289 (2006)

An Improved Particle Swarm Optimization for Complex Optimization Problems

Kezong Tang^{1,2}, Binxiang Liu¹, and Jia Zhao²

¹ Jingdezhen Ceramic Institute Information Engineering Institute,
Jingdezhen Jiangxi 333403, China
tangkezong@126.com

² School of Information Engineering, Nanchang Institute of Technology,
Nanchang 330099, China
zhaojia925@163.com

Abstract. An improved particle swarm optimization (IPSO) is proposed where a general center particle is incorporated into particle swarm optimization (PSO) with linearly decreasing inertia weight factor in this paper. The general center particle is formed by the center of the best-found positions of all particles in IPSO. It has potential capacity to get good positions and guide the search direction of the whole swarm because of frequently appearance as the best particle of the swarm. Numerical results and comparison on a set of benchmark optimization functions show the proposed algorithm is a promising optimization method in obtaining better solutions.

Keywords: Particle swarm optimization, General Center particle, Optimization problem.

1 Introduction

Particle swarm optimization (PSO) is a population-based stochastic global optimization method, firstly introduced by Kennedy and Eberhart [1]. In the past ten years, many researchers have devoted to further enhancing its performance in various ways and applied its improved variations to many areas[2-4].

Among the existing improved approaches, many methods and strategies have been proposed to enhance the performance of PSO, such as linearly decreasing weight particle swarm optimization by Shi and Eberhart [5], constriction PSO by Clerc [6], concurrent PSO by Baskar and Suganthan [7], attractive and repulsive PSO by Riget [8] and CenterPSO by Liu [9]. In spite of their diversity, but all of them still abide by the same principle of swarm intelligence. During convergence, all particles are attracted by the global best-found position by the whole swarm (gbest), and it directly affects the convergence speed and also the quality of the final solution. Therefore, the improvement of the gbest is a very meaningful work for the effectiveness and efficiency of the PSO.

In this paper, a general center particle (GCP), whose position is updated with the center of all best-found positions of particles at every iteration, is incorporated into the PSO algorithm with linearly decreasing inertia weight factor. The proposed method,

called, IPSO, is based on the common features the same as the traditional PSO. In IPSO, due to the GCP frequently appears in the form of global best particle of the swarm, so it has significantly potential to guide the search process of the whole swarm.

The remaining content of this paper is organized as follows: In section 2, traditional PSO and its optimization mechanism is described, and the general center particle on the performance of IPSO is discussed in detail. In section 3, the IPSO and CenterPSO are compared on a set of benchmark function optimizations. Finally, conclusions and directions of future work are given in section 4.

2 Traditional PSO

In the traditional PSO algorithm, Every particle updates own position and velocity according to the following equations:

$$V_{id}^{t+1} = w \cdot V_{id}^t + c_1 \cdot r_1 \cdot (P_{id} - X_{id}^t) + c_2 \cdot r_2 \cdot (p_{gd} - X_{id}^t) \quad (1)$$

$$X_{id}^{t+1} = X_{id}^t + V_{id}^{t+1} \quad (2)$$

where V_{id}^t and X_{id}^t represent the velocity and position of the i th dimension of particle i in the solution space at t th iteration step respectively. c_1 and c_2 are two positive constants in the range of 0-2. P_{id} denotes the local position found by particle i within t iterations. p_{gd} denotes the best position found among all the particles through the objective function. r_1 and r_2 are two independent random numbers uniformly distributed in the range of [0,1]; $w \in (0,1)$, an inertia weight factor, can control the amount of recurrence in the particle's velocity so as to maintain the balance between global and local search.

3 The Proposed Algorithm

The global best-found position by the whole swarm ($gbest$) is an extremely important position which guides the swarm to the final best position. So the $gbest$ plays an extremely important role in supplying significantly information for capturing the final best position ($fbest$). In the analysis on optimization mechanism of PSO, it is noticeable that, the $gbest$ is to approach the $fbest$ of solving optimization problem as iteration proceeds, but it is not exactly located in the $fbest$. Therefore, the $gbest$ is a crucial position during search because it can affect the convergence speed and also the quality of the final solution. To some extent, the improvement of $gbest$ depends on that of $pbest$ of each particle in the swarm. While flying through the search space, all particles move in the certain region where $gbest$ locates. Then, their positions are usually different and approximately around $gbest$. Meanwhile, the current $pbest$ of each particle is also distributed around the $gbest$ for reasons of many stochastic factors, and the center of all particles' $pbest$ is very promising to approach the $fbest$ further, and in many cases, this center is closer to the $fbest$ during the search than the $gbest$. For convenient observation, two-dimensional Rastrigin function optimization is illustrated. Figs.1 (a)-(f) show the distribution of the particles at the first, 10th, 30th, 50th, 80th, 100th iteration when

two-dimensional Rastrigin function was optimized with traditional PSO with 10 particles. The signs: hollow point, circle, star and triangle denote the positions of the particle, f_{best} , center of all particles' p_{best} , g_{best} , respectively. As shown in these figures, the following phenomenon can be seen. The region of swarm activity is constantly changing. Both g_{best} and center of all particles' p_{best} are increasingly approaching the f_{best} as search proceeds, but at a given time, center of all particles' p_{best} is closer to the f_{best} during the search than the g_{best} . In a sense, if this center replaces the g_{best} as the current global best position while this center is superior to the g_{best} , then it can guide the whole swarm to f_{best} and accelerate convergence. To sum up, the center of all particles' p_{best} is an extremely important position and eventually all the particles statistically shrinks to it through the iteration. However, this center position is not explicitly considered in the traditional PSO algorithm and its other improved variations.

In the paper, a general center particle is introduced to the tradition PSO algorithm with the aim to explicitly visit this center position during search. After $N-1$ particles update their positions as the usual PSO algorithms at the t th iteration, a general center particle is updated according to the following formula:

$$X_{cd}^t = \frac{1}{N-1} \left(\sum_{i=1}^{N-1} P_{id}^t \right) \quad (3)$$

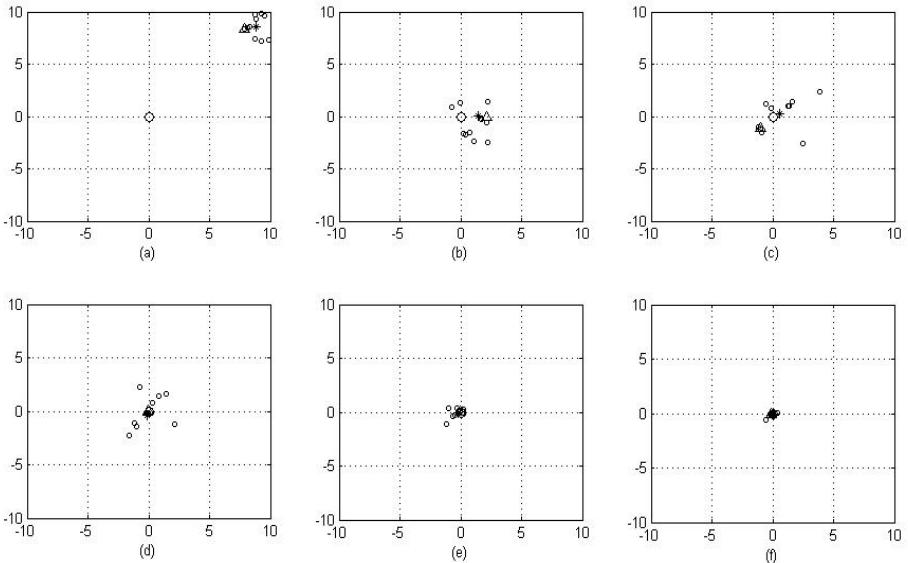


Fig. 1. The distribution of particles: (1) at the first iteration; (b) at the 10th iteration; (c) at the 30th iteration; (d) at the 50th iteration; (e) at the 80th iteration; (f) at the 100th iteration.

The general center particle is involved in the common operations the same as the ordinary particle except for the velocity calculation, such as the survival as a member of swarm by coexistence and cooperation, individual contribution to the best position of

the swarm and fitness evaluation. Although it is only one particle, it imposes great effect on the swarm. The pseudo-code of IPSO is given as follows:

- Step1.** Initialize randomly all particles positions X_i^t and velocities V_i^t .
- Step2.** Evaluate fitness value as $f(X_i^t)$ according to the optimization problem.
- Step3.** Assign best positions $P_i^t = X_i^t$ with $f(P_i^t) = f(X_i^t), i = 1, \dots, P$.
- Step4.** Update the position of general center particle X_c^t according to Equation (3).
- Step5.** Find g_{best} according to $f(P_g^t) = \min\{f(P_1^t), f(P_2^t), \dots, f(P_{N-1}^t), f(X_c^t)\}$
- Step6.** Terminate if P_g meets problem requirements.

4 Experimentation

We perform experimentation on a set of well-known benchmark optimization problems for comparison with CenterPSO in the literature [9]. This paper utilizes the benchmark function set, shown as follows.

$$f_1(x) = \sum_{i=1}^{n-1} (100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2) \quad (-100 \leq x_i \leq 100), \text{ Rosenbrock function.}$$

$$f_2(x) = \sum_{i=1}^n (x_i^2 - 10 \cos(2\pi x_i) + 10) \quad (-10 \leq x_i \leq 10), \text{ Rastrigin function.}$$

$$f_3(x) = \frac{1}{4000} \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1 \quad (-600 \leq x_i \leq 600), \text{ Griewank function}$$

The same set of parameters was assigned for two algorithms: inertia weight factor w linearly decreased from 1 to 0.4; both learning coefficient were $c_1=c_2=2$; V_{\min} was equal to $-V_{\max}$; To investigate the effect of high dimension on searching equality of the proposed IPSO method, for all functions three different dimensions ($D=10, 20, 30$) were tested; Correspondingly, maximum numbers of iteration were set to 1000, 1500 and 2000, respectively. Four population sizes ($Popsize=20, 40, 80, 160$) were used for each function with different dimensions, respectively. The IPSO includes one general center particle and $N-1$ ordinary particles the same as the CenterPSO. In order to eliminate stochastic discrepancy, we perform 100 runs for each experimental setting. The search domain, asymmetric initialization range and V_{\max} were listed in Table 1.

Table 1. Parameter settings for search domain, asymmetric initialization range and V_{\max}

Function	Search domain	Asymmetric initialization range	V_{\max}
f_1	[100,100]	(15,30) ⁿ	50
f_2	[-10,10]	(2.56, 5.12) ⁿ	5
f_3	[-600,600]	(300,600) ⁿ	300

Table 2. Testing results for the Rosenbrock function

Size	Dim	Max iteration	IPSO		CenterPSO	
			Mean	SV	Mean	SV
20	10	1000	17.9873	50.1838	18.4211	55.3924
	20	1500	71.4491	84.3241	72.5825	85.3435
	30	2000	91.7244	102.9559	106.0593	113.3050
40	10	1000	3.9573	9.1311	5.6648	16.1659
	20	1500	30.8986	30.1508	34.7552	53.7457
	30	2000	50.9993	44.8715	50.3010	51.7102

Table 3. Testing results for the generalized Rastrigin function

Size	Dim	Max iteration	IPSO		CenterPSO	
			Mean	SV	Mean	SV
20	10	1000	6.1185	2.2537	8.4523	3.6961
	20	1500	29.0228	24.2395	30.5974	8.4226
	30	2000	49.7568	39.8844	59.0294	16.1252
40	10	1000	6.0394	2.4217	6.7658	2.7349
	20	1500	22.1582	18.2981	23.2551	6.7610
	30	2000	40.1857	24.7975	43.4639	11.4669

Table 4. Testing results for the generalized Griewank function

Size	Dim	Max iteration	IPSO		CenterPSO	
			Mean	SV	Mean	SV
20	10	1000	0.1274	0.0677	0.1378	0.0701
	20	1500	0.2442	0.1010	0.2903	0.1196
	30	2000	0.5808	0.1444	0.6320	0.1267
40	10	1000	0.1168	0.0576	0.1187	0.0659
	20	1500	0.0441	0.0294	0.0481	0.0309
	30	2000	0.1718	0.0686	0.2079	0.0766

Tables 2-4, respectively, listed the mean fitness value (Mean) and standard variation (SV) of the best solutions averaged over 100 runs on five benchmark functions with each experimental setting. As shown in these tables, For functions f_1 , f_2 and f_3 , the proposed algorithm outperforms the CenterPSO algorithm in terms of Mean and SV.

Particularly for f_3 , IPSO shows a strong ability to achieve the global optimum because its solution distribution at every run is focused on the final best solution. For function f_2 , the testing results of IPSO are a bit close to that of CenterPSO, and both algorithms can obtain the small fitness value and standard variation. However, f_2 and f_3 have little linkages among different dimensional values. It means the proposed IPSO can improve the performance of algorithm than CenterPSO at least above 20%.

To sum up, in view of the algorithm efficiency and effectiveness in term of statistical data of testing functions, it can be anticipated that IPSO method remains quite competitive as compared to CenterPSO. We can draw the following conclusions: IPSO is fit for solving high dimensional optimization problems with strong linkages, especially for dimensionality is larger than 10. This further indicates the effectiveness of the center particle introduced in traditional PSO.

5 Summary

In this paper, we proposed general center PSO algorithm where a general center particle was introduced into traditional PSO with linearly decreasing inertia weight factor. The general center particle usually gets good position and often appears in the form of the best particle with $gbest$. Therefore, it has greatly ability to guide the search direction of the whole swarm and accelerate convergence speed. Compared with CenterPSO algorithm, the IPSO has obvious superiority in terms of the mean fitness and standard variation. More tests on different optimization problems will be one part of our future work.

Acknowledgement. This paper is supported by the National Natural Science Foundation of China (Grant No. 61202313, 61202318, 61261027, 31260273), the national science and Technology Support Program Fund Project of China (2012BAH25F02) and The Natural Science Foundation of Jiangxi Province of China (GJJ12642, GJJ12514, 2012BAB201044) and Technology Project of provincial University of Fujian Province (JK2011040).

References

- [1] Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceeding of the IEEE International Conference on Neural Network, Perth, Australia (1995)
- [2] Fan, S.-K.S., Liang, Y.-C., Zahara, E.: Hybrid simplex search and particle swarm optimization for the global optimization of multimodal functions. Engineering Optimization 36(4), 401–418 (2004)
- [3] Bergh, F., Engelbrecht, A.P.: Training product unit networks using cooperative particle swarm optimizers. In: Proceedings of International Joint Conference on Neural Network, vol. 1, pp. 126–131 (2001)
- [4] Zahiri, S.H., Seyedin, S.A.: Swarm intelligence based classifiers. Journal of the Franklin institute 344(5), 362–376 (2007)

- [5] Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: Proceedings of the IEEE Conference on Evolutionary Computation, Anchorage, AK, USA (1998)
- [6] Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation 6(1), 58–73 (2002)
- [7] Baskar, S., Suganthan, P.: A novel concurrent particle swarm optimization. Proceedings of the Congress on Evolutionary Computation 1, 792–796 (2004)
- [8] Riget, J., Vesterstrøm, J.S.: A diversity-guided particle swarm optimizer-the ARPSO. Technical Report 2002-02, EVALife, Department of Computer Science, University of Aarhus (2002)
- [9] Liu, Y., Qin, Z., Shi, Z.W., Lu, J.: Center particle swarm optimization. Neurocomputing 70(4-6), 672–679 (2007)

An Improved Method for Oriented Chamfer Matching

Jian Dong^{*}, Changyin Sun, and Wankou Yang

School of Automation, Southeast University, Nanjing 210096, China
dongjian72@163.com

Abstract. Oriented Chamfer Matching is much more tolerable to detect object in images with clustered background. However, it is still more reliable on object template. Focused on this problem, this paper proposed a method flexibly creating deformable templates. The multiple deformable templates were created through simulating the 3D perspective projection of 2D template. The method called Procrustes Alignment was utilized to combine all the deformable templates into a unified one. The proposed method was tested on ETHZ shape class dataset and better detection results were acquired.

Keywords: Oriented Chamfer Matching, Clustered Background, 3D Perspective Projection, Procrustes Alignment.

1 Introduction

Object detection and matching are key problems of computer vision, scene understanding, image analysis and other topics. However, due to the existing of changes in scale, illumination, and viewpoint and so on, the detection problem has become intractable. Human can easily recognize an object according to its contour, even part of the contour, which means that the contour can provide effective information for object recognition. Meanwhile, the contour-based or edge-based algorithm has strong adaptability to illumination changes, which makes the algorithm more effective and applicable. As one of the most important contour-based or edge-based algorithm, Chamfer distance transform (Chamfer DT) has been widely used for multi-classes object detection and matching. And recent improvements have made the algorithm more efficient and suitable for real-time detection.

This paper based on the Chamfer DT algorithm combining orientation information, i.e. oriented chamfer matching (OCM), in images with cluttered background. A flexible way to create object template is introduced, the detection effect is greatly improved combining OCM algorithm.

2 Related Work

There are lots of literatures using Chamfer DT in computer vision. Chamfer distance was first introduced by Barrow et al. [1] in 1977 with the goal of matching two sets of

* Corresponding author.

edge points and was greatly improved by G. Borgefors [2] in 1988 through hierarchical structure. In [3], Chamfer distance was used to make second detection of possible human position in the image and the best pose of human was detected. Abdul Ghafoor et al.[6] proposed modified chamfer matching algorithm (MCMA) achieving simple detection of airplane model. Qiang Zhang et al. [4] based on the hierarchical chamfer matching and proposed fast distance image pyramid method.

However, due to the inherent drawbacks of chamfer matching, it is not effective when detecting objects in images with clutter background. Originally, the template and query image edges were quantized into discrete orientation channels and individual matching scores across channels were summed [10, 11]. In [9], Ming-Yu Liu used linear representation to represent the template model for simplicity, gave a directional chamfer matching (DCM) score as search criterion and proposed three-dimensional distance transform to compute the matching cost in linear time. Jamie Shotton et al. [5] used Chamfer matching combining edge orientation information and built a codebook of fragments realizing the multi-scale and multi-class visual object recognition. Based on oriented chamfer matching, Jamie Shotton et al. [7] presented a novel categorical object detection scheme that used only local contour-based features. Tianyang Ma et al. [8] based the method that if the part of an image matches well both with the template and some random shapes, it is more likely that this part is more likely to contain clustered background, rather than the object and proposed some normalizer to auxiliary detect the object..

3 Chamfer Matching

Chamfer matching was first proposed to search the best fitting of points from two different sets and in image. Let $U = \{u_i\}$ and $V = \{v_i\}$ be the two edge points sets indicating the template edge points set and query image edge points set respectively. Chamfer distance between points sets U and V is the average of distances between each point $u_i \in U$ and its nearest edge point in V . Considering the translation and rotation parameter, the score of OCD can be written as:

$$\bar{d}_{OCD}(T(U; s), V) = (1 - \lambda)\bar{d}_{CM}(T(U; s), V) + \lambda d_{oreint}(T(U; s), V) \quad (1)$$

Where $\bar{d}_{CM}(U, V) = \frac{1}{|U|} \sum_{u_i \in U} \min_{v_j \in V} (\min \|u_i - v_j\|, \tau)$ denotes the original distance transform combining a truncation parameter, $T(U; s)$ is the points set of points set U after translation according to parameter $s = (t_x, t_y, \theta)$, $|U|$ is the number of points in U ; $d_{oreint}(U, V) = \frac{2}{\pi |U|} \sum_{u_i \in U} |\phi(u_i) - \phi(ADT(u_i))|$, and $\phi(u_i)$ denotes tangent direction at point u_i ranging between zero and π , $ADT(u_i) = \arg \min_{u_j \in U} \|u_i - v_j\|$, $v_j \in V$ specifies the nearest point.

4 Create Flexible Deformable Template

In oriented chamfer matching, the detecting results are more reliable on the template. In this part, the method that simulates the perspective projection of template was introduced and multiple deformable templates were created. A method called Procrustes Alignment was introduced to combine the deformable templates into a unified one.

4.1 Simulation of Perspective Projection

The 3D perspective projection is as follows. Let (x, y, z) to be the point in 3D space and (X, Y, Z) is the projection point.

$$\begin{cases} X = x / (px + qy + rz + 1) \\ Y = y / (px + qy + rz + 1) \\ Z = z / (px + qy + rz + 1) \end{cases} \quad (2)$$

where, p, q, r are the projection parameter.

The 3D perspective projection is simulated from original 2D points through manually setting the z coordinate according to x sequence by a space 0.1. So the z coordinate is supplemented. The 6 deformable templates are shown in Fig.1.

4.2 Procrustes Alignment

Procrustes analysis computes the best set of transformations that relate matched shape data [12]. Based on the Procrustes theory, some of the similar shapes can be combined into a unified one. Let $M = [x_1, \dots, x_k; y_1, \dots, y_k; z_1, \dots, z_k]$ be the data matrix of template edge points. Executing SVD decomposition on $M_1^T M_2 = UDV^T$, the rotation matrix $\Gamma = UV$ can be got. The Procrustes distance is defined as follows:

$$d(M_1, M_2) = \|M_2 - \beta M_1 \Gamma\| \quad (3)$$

The Procrustes alignment algorithm is realized as the following procedures:

- (a). Choose an arbitrary shape as the mean shape M_0
- (b). Align the rest of the shapes to M_0 : $M_i^p = M_i \Gamma_i$
- (c). Set new M_0 to the average of the Procrustes-aligned shapes: $M_0 = \frac{1}{n} \sum_{i=1}^n M_i^p$
- (d). Repeat step (b) and (c) until M_0 is stabilized

The original template, 6 deformable templates and the unified template are shown in Fig.1.

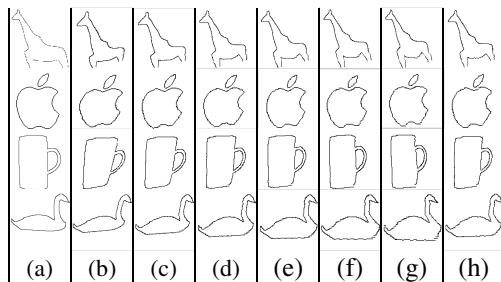


Fig. 1. The original template, 6 deformable templates and the unified template. (a) Original template, (b)~(g) 6 deformable templates, (h) unified template by Procrustes alignment.

5 Experiments

In this part, the proposed algorithm was tested in ETHZ shape class dataset. The proposed unified model has stronger adaptability than the algorithm in [9]. If simultaneously using some of the deformable templates to detect images in the dataset, the adaptability will be much stronger. The test results are shown in Fig.2.

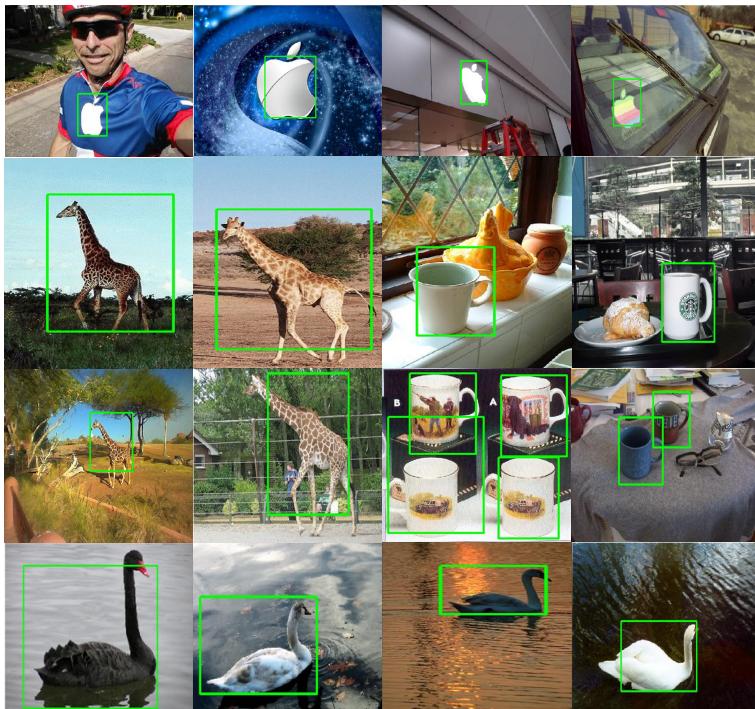


Fig. 2. The test results on ETHZ shape class dataset

6 Conclusion

Based on the problem of Chamfer matching is reliable on the form of template. This paper gives a flexible way to create deformable templates which are more useful to detection objects in ETHZ shape class dataset. By manually setting the value of z coordinate according to x sequence, the 2D template edge points were changed into 3D. According to Procrustes distance and Procrustes Alignment, all the deformable templates were combined into a unified one which is more effective. The detection results on the images in ETHZ dataset show the effectiveness of the algorithm.

Acknowledgements. This work is supported by National Natural Science Foundation of China (No. 61273023) and the Ph.D. Programs Foundation of Ministry of Education of China (No. 20120092110024).

References

1. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. In: Proc. 5th Int. Joint Conf. Artificial Intelligence, pp. 659–663 (1977)
2. Borgefors, G.: Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. Pattern Analysis and Machine Intelligence* 10(6), 849–865 (1988)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. *International Journal of Computer Vision* 61(1), 55–79 (2005)
4. Zhang, Q., Xu, P., Li, W., Wu, Z., Zhou, M.: Efficient edge matching using improved hierarchical chamfer matching. In: IEEE International Symposium on Circuits and Systems, pp. 1645–1648 (May 2009)
5. Shotton, J., Blake, A., Cipolla, R.: Multi-scale categorical object recognition using contour fragments. *IEEE Trans. Pattern Analysis and Machine Intelligence* 30(7), 1270–1281 (2008)
6. Ghafoor, A., Iqbal, R.N., Khan, S.: Image Matching Using Distance Transform. In: 13th Scandinavian Conference on Image Analysis, pp. 654–660 (June 2003)
7. Shotton, J., Blake, A., Cipolla, R.: Contour-Based Learning for Object Detection. In: IEEE International Conference on Computer Vision, pp. 503–510 (October 2005)
8. Ma, T., Yang, X., Latecki, L.J.: Boosting Chamfer Matching by Learning Chamfer Distance Normalization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 450–463. Springer, Heidelberg (2010)
9. Liu, M.-Y., Tuzel, O., Veeraraghavan, A., Chellappa, R.: Fast Directional Chamfer Matching. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1696–1703 (June 2010)
10. Danielsson, O., Carlsson, S., Sullivan, J.: Automatic learning and extraction of multi-local features. In: 12th International Conference on Computer Vision, pp. 917–924 (September 2009)
11. Gavrila, D.M.: Multi-feature hierarchical template matching using distance transforms. In: International Conference on Pattern Recognition, pp. 439–444 (August 1998)
12. Adrien, B., Daniel, P., Marco, L.: Stratified generalized procrustes analysis. *International Journal of Computer Vision* 101(2), 227–253 (2013)

Computer Vision Based Pose Bias Detection of Shield Tunneling Machine

Jiannan Chi¹, Lei Liu¹, Jiwei Liu², Changyin Sun¹, and Weiping Zhang¹

¹ School of Automation and Electrical Engineering, University of Science and Technology, Beijing, 100083, China

² School of Automation, Shenyang Aerospace University, Shenyang, 110136, China

Abstract. Shield tunneling machine is a kind of special engineering machinery applied to tunnel excavation. In the process of construction, shield tunneling machine maybe produce azimuth deviation, which influences the direction of tunnel excavation. Artificial laser spot observation method is commonly used in shield deviation correction of engineering. However, this method will increase the burden of operators and is prone to misjudge the deviation. For the practical application, this paper put forward a deviation detection method of shield posture based on computer vision. In this method, the fixed laser light source produces light spot on the chessboard target which is installed in the rear of shield machine. The image of chessboard is extracted through the camera. The markers and laser spot in the image is detected by using image processing and analysis method. Which can reflect the deviation in the process of advance of shield machine. Experimental results have verified the validity of this method.

Keywords: Image processing, Hough transform, adaptive threshold, Shield machine rectification.

1 Introduction

Shield machine is a kind of tunneling engineering machinery, which is widely used in subway construction, railway, highway engineering and so on [1]. In practical engineering application there exist many factors affecting the process of excavation, which cause that shield tunneling deviate from the advance direction. Therefore real-time detection of position and posture of shield tunneling machine during construction process is important for the posture control [2].

Methods of shield posture measurement commonly used in engineering mainly include manual measurement method, such as Scale Method [3], Three-point Method [4], and automatic measurement method, such as Gyroscope method and Automatic Total Station method [3,5,6,7,8].

The methods of manual measurement and automatic measurement have been widely used in shield machine gesture measuring [3]. In the practical application, there is a method known as artificial laser spot observation as shown in Fig.1. A chessboard target plate is set up in the rear of shield machine. A fixed laser source is installed far behind the shield machine. A laser spot on the surface of chessboard is generated when emitting laser beam is at chessboard. With the movement of the shield machine the location of laser spot on chessboard changes. Posture deviation of shield machine can

be judged from the displacement of laser spot. In this paper we use computer vision to improve the method of artificial laser spot observation. The simulation experiment results demonstrated the effectiveness of the method.

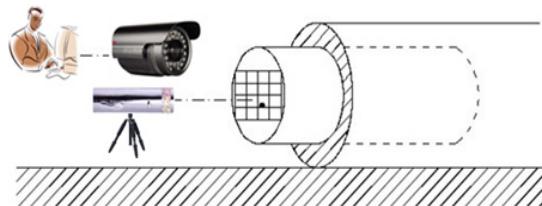


Fig. 1. Sketch map of artificial laser spot offset method

2 Introduction of Shield Machine Correction Method Based on Laser Irradiation

Computer vision based improved artificial laser spot observation method is as follows. Firstly, four circle Retro-reflective markers are set up on the corners of chessboard target, and in front of the chessboard a camera with coaxial light source is placed for collecting chessboard target and laser spot images. Through image processing and analyzing the laser spot and markers are segmented in the image and their coordinates are obtained. Therefore deviation of shield machine can be estimated via the center coordinates of laser spot and markers.

There are two methods that can be used to measure the deviation of shield machine. One is to make the camera fixed while shield machine move forward. Another is to make the camera move along with the shield machine.

(1) The camera is fixed in front of chessboard target. With LED coaxial light shining on the Retro-reflective markers distributed on the chessboard target corners, regions of Retro-reflective markers in the image become particularly bright. In every moment the location of markers is detected and the displacement of marker can be calculated to reflect the deviation of shield machine.

(2) Another deviation detection mode is that the camera is moving along with shield machine. The relative position between camera and chessboard doesn't change. Location of the markers in the image doesn't change as well. In this case deviation of shield machine is reflected by the location of laser spot on chessboard.

3 Deviation Detection Method of Shield Machine

3.1 Deviation Detection in Case of Fixed Camera

As the spatial location of four markers on chessboard corner are fixed and known, they work as measuring points as well as calibration points, which are used for calibrating proportional coefficient between image scale and the actual dimension. According to the image coordinates of four markers and the proportional coefficient, actual offset of

the four markers can be obtained. The real offset and the rotation angle of shield machine can be determined. The steps are as follows:

(1) In order to detect deviation we must locate the markers in image. Because of the markers on the chessboard, in the initial operation chessboard region in image need to be determined. At same time, the initial position of laser spot in the image also can be determined by interaction. The initial location of laser spot is the initial reference point. Through mouse operation we can determine chessboard region in image and record the length and width of chessboard (dx, dy). We can also determine the center coordinate of laser spot which is recorded as (x_j, y_j) by mouse operation.

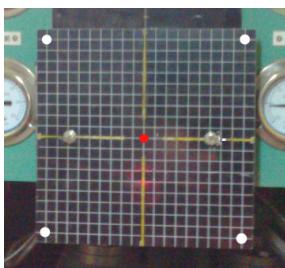


Fig. 2. Chessboard



Fig. 3. Binarization Result

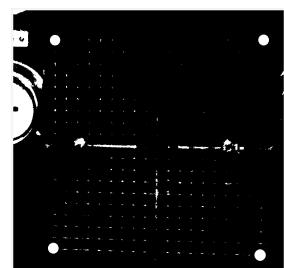


Fig. 4. Opening operation Result

(2) To segment the marker targets in image under different lighting condition . Adaptive threshold method is used to get binary image. Procedure is as follows:

A, Image conversion: Convert color images to grayscale images.

B, Image binarization based on adaptive threshold: calculate the histogram of grayscale image to obtain the maximum of gray value. This grey value is precisely the background grayscale, which is recorded and used as an image-binary threshold. After binarization, chessboard background is restrained. Those regions with high grey-scale will be retained, which may be marker area.

C, Remove small noise and retain high gray targets, pseudo targets can be removed further. Results of image binarization and operation are shown in figure 3 and 4.

(3) As the advance of shield machine, the camera is fixed, so chessboard target area in the image will get smaller. The chessboard target area determined by frame drawing will inevitably contain some background objects such as analog measuring instruments as shown in figure 2. So after image binarization introduced above in (2), some of the larger high gray masses are bound to be left in the image. To find the circular point target, and remove other shapes or larger circular mark (such as the dial, etc.), a Hough transform circle fitting method based on adaptive threshold is put forward in this paper. Specific steps are as follows:

A, As the mark points are quite small in the image, in order to find them accurately, the initial accumulator threshold of Hough transform should be set to a larger value at first (the value is 200). Then use Hough transform for fitting circles and find the center of the circles. Coordinates (x, y) and radius r of the circles are recorded in set A.

B, If the elements in set A were less than the number of target points (the number is 4 in this paper), repeat step (1) after subtracting 10 from the initial accumulator threshold value mentioned above. Otherwise go to the next step.

C, According to the formula $K = y_2 - y_1 / x_2 - x_1$, calculate the slope of connecting lines between the center of the circles in turn.

D, If $K \approx 0$ or $1 / K \approx 0$, record coordinates of the two centre points corresponding to K in set B.

E, Remove the two points in step D from set A and repeat steps B, C.

F, If the elements in set B were less than the number of target points 4, empty set B and repeat step A to E after subtracting 10 from the initial accumulator threshold value in step (1). Otherwise, go to the next step.

G, Sequence elements from the smallest to the largest according to the radius size in set B, and record the first four in set C.

(4) Calculate the sum of squares of the center coordinates $x^2 + y^2$ in set C. Determine the mark point in upper left, upper right, lower left and lower right corner according to relationship between horizontal and vertical coordinates, record as $NUM1, NUM2, NUM3$.

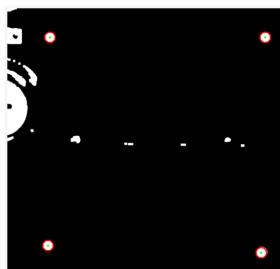


Fig. 5. The final result of mark points fitting by Hough transform

(5) As is known, the actual distance in transverse and longitudinal direction between mark points is 18 grids. According to the coordinates of laser spots and mark points, laser spot location can be obtained. Horizontal direction is as follows:

$$C_x = \frac{18(x_j - num1.x)}{(num4.x - num1.x) - 9} \quad (1)$$

Vertical direction is as follows:

$$C_y = \frac{-18(y_j - num1.y)}{(num4.y - num1.y) - 9} \quad (2)$$

Since actual distance between mark points is preset, according to the image coordinates, the ratio between actual distance and image distance can be identified. Shield machine's actual offset can be calculated on the basis of the mark points' offset and the shield machine's rotation angle can also be calculated through the same way.

3.2 Deviation Detection with Camera Following Shield Machine

When camera follows the movement of shield machine, posture of the shield machine can be reflected by the location of the laser spot on the chessboard. Calibrating the proportional coefficient between the image scale and actual scale is necessary in the initial image process by reason that four mark points' positions in image remain unchanged. Determine the red laser point coordinates in the image, according to the changes of coordinates and proportional coefficient, offset of shield machine is obtained. The image processing steps are as follows:

(1) Because the red laser spot will not deviate from the chessboard area, then capture images after removing the mark points and extract pixel values of R channel. Save the result as a new image shown in figure 6.

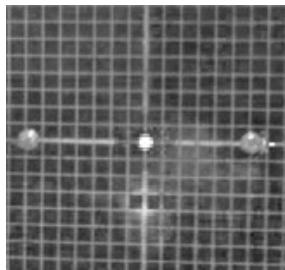


Fig. 6. Laser spot areas R channel schematic diagram

(2) Because shape and brightness of laser spot will be constantly changing, it is reasonable to adopt the method of finding connected domain for laser spot location. OTSU method is used for image binarization and subsequently searching for connected domain. The next is to find out the largest connected domain and calculate its bar centric coordinates, which are considered to be red laser point location in the image. Mark the coordinates as (x_j, y_j) .

(3) Calculate the difference between two different measurement results of laser spot location coordinates in the image. Divided by dx which is obtained by calibration, the moving grid number in horizontal direction is acquired. Similarly, divided by dy , the moving grid number in vertical direction can be obtained.

4 Experimental Results

In order to verify the effectiveness of the method, mark points and red laser spot are added for simulation on the collected images of actual chessboard target. Using the proposed mark points and red laser point detection method, experiments are conducted in PC assembled with dual-core 3.3 GHz CPU, Windows 7 and memory size is 4 GB. The image resolution provided by digital camera is 2592*1944 pixels. Although mark points and laser spot of different brightness are applied in simulation tests, the proposed mark points detection method and laser point detection method can stably detect and locate mark points and laser spot.

From figure 7 we can see, the adaptive threshold target segmentation method can remove chessboard target background and reserve mark point targets even under different illumination conditions. Circle fitting method based on adaptive threshold Hough transform can make a good fitting of mark points and eliminate circular non-marked point targets. The same is shown in figure 8, the adaptive threshold segmentation method based on color histogram can also divide laser spot target well.

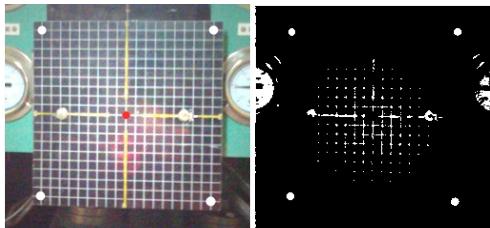


Fig. 7. Threshold segmentation results

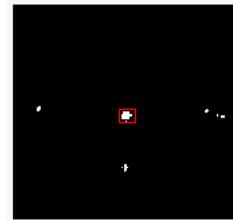


Fig. 8. Red laser spot segmentation result

Laser spot center test results in different locations are shown in table 1.

Table 1. Measurement results in different laser spot locations

actual value in horizontal direction	measured value in horizontal direction	absolute error	relative error (%)
2	2.012	0.012	0.600
1.5	1.514	0.014	0.933
1	0.992	0.008	0.800
actual value in vertical direction	measured value in vertical direction	absolute error	relative error (%)
2.5	2.493	0.007	0.280
2	1.981	0.019	0.950
0.5	0.507	0.007	1.400

From the experimental results we can learn, the algorithm can accurately detect the location of the mark points and laser spot under different illumination.

5 Conclusion

Shield machine rectifying method based on artificial laser spot observation has many deficiencies, such as increase burden of operator and easy to misjudge. Geared to the needs of actual engineering requirements, a kind of shield machine offset detection method based on computer vision technology is put forward in this paper. In this method fixed laser light source is used to irradiate chessboard target to produce light spot and through the camera chessboard target images can be obtained. When the camera is fixed, by segmenting the mark points on the chessboard target and detecting their coordinates in the image, the offset of shield machine can be calculated. When the

camera moving with shield machine, first the light spot position where shield machine starts driving is marked as initial reference position, then deviation value between laser spot trajectory and original location can be detected and shield machine's offset in forward movement can be further reflected. The experimental results have verified the effectiveness of the image processing and analysis method in this paper.

Acknowledgement. Supported by the Opening Project of Key Laboratory of operation safety technology on transport vehicles, Ministry of Transport, PRC.

References

1. Dong, W.-D., Ren, G., Ma, L.: The Principle of Laser Navigation System of The Tunnel Boring Machine. *Engineering of Surveying and Mapping* 14(4), 61–64 (2005)
2. Tang, Z., Zhao, J., Peng, G.: Measurement and Calculation Method for Real-time Attitude of Tunnel Boring Machine. *China Civil Engineering Journal* 40(11), 92–98 (2007)
3. Ouyang, P., Wu, B.-P.: Comparison of the Precision of Several Shield Posture Measurement Techniques. *Chinese Journal of Engineering Geophysics* 3(4), 304–310 (2006)
4. Xu, Z., Limin, Z.: Optimal Selection of Attitude Control Points of Shield Machine. *China Mechanical Engineering* 20(8), 902–907 (2009)
5. Huang, X.-B., Ou, Z.-M., Zhang, Y.-C., Jiang, Y.-M.: Manual measurement principle of tunnel boring machine attitude. *Science of Surveying and Mapping* 36(3), 25–29 (2011)
6. Pan, M.-H., Zhu, G.-l.: Study of Measure Methods of the Automatic Guiding System of Shield Machine. *Construction Technology* 34(6), 34–37 (2005)
7. Wang, C.-L., Zhang, Y.-C.: Study on Real-time Attitude Survey of Shield Machine for Metro Works. *Tunnel Construction* 27(6), 33–35 (2007)
8. Liu, G.: The research of the measurement technology about the shield posture parameters. *Technological Development of Enterprise* 26(4), 28–32 (2007)

Integrative Hypothesis Test and A5 Formulation: Sample Pairing Delta, Case Control Study, and Boundary Based Statistics

Lei Xu

Department of Computer Science and Engineering
Chinese University of Hong Kong
lxu@cse.cuhk.edu.hk

Abstract. This paper continues the previous preliminary study on integrative hypothesis test (IHT) (Xu, LNCS7751, 2013). First, the coverage of IHT studies are elaborated from four aspects. Then, the previous preliminary A5 formulation for IHT is developed into one that integrates multiple individual tasks of discriminative analysis to improve hypothesis test with enhanced reliability. Next, a sample-pairing-delta based nonparametric statistics is proposed and its application to case control study is addressed. Moreover, a parametric separating boundary is further embedded into hypothesis test with statistics based on how far samples are away from the boundary, under which sample classification and hypothesis testing are coordinately implemented.

Keywords: Integrative hypothesis test, A5 formulation, False discovery rate, Accumulated reliability, Sample pairing delta, case control study, SNP analysis, boundary based statistics.

1 Introduction

Informally, the term *integrative hypothesis testing* (IHT) is used in [1] for discriminative analysis to integrate the evidences associated with structured variables on which one *hypothesis* is supported and to integrate the outcomes of many different hypothesis tests in order to reach a final conclusion. Moreover, taking analyses of gene expression and exome sequencing as examples, one so-called Geno-Pheno A5 analyzer is proposed in [1] to apply an A5 formulation of the problem solving paradigm [2], resulting in a general procedure for IHT implementation.

Many applications in bioinformatics and other tasks of big data analysis involve discriminative analyses on samples from different populations. Without losing much generality, this paper focuses on such tasks of two populations, which is featured by subtasks viewed from three complementary perspectives. One aims at classifying samples into their corresponding populations by a discriminant rule, which is usually named classification or decision and widely encountered in the literatures of pattern recognition and machine learning. Being popular in the literature of bioinformatics and medical informatics, the second is made under the name of *hypothesis test*, which evaluates whether two populations of samples are significantly different according to

some discriminative statistics. Both the two subtasks are made on a finite set of samples and thus highly depend on another subtask called *feature or variable selection*. The selected features or variables form the domain on which each of the two subtasks is performed. The two subtasks are related but implemented subject to different performance measures that are not monotonically related. Thus, one best set of selected features for one subtask may not be necessarily one best set for the other.

This paper continues the IHT study made in [1], starting at elaborating the coverage of IHT studies from the following aspects:

- How to integrate information associated with multiple features for testing an overall *hypothesis*. Typical topics include how to develop an overall statistics, e.g., a general choice is suggested by Eq.(24) in [1] with samples of structured features expressed in matrix format, and how to compute statistics efficiently, e.g., as partly discussed in Sect.4.1 of [1], as well as how to estimate the p-value and q-value.
- How to integrate the outcomes of multiple individual hypothesis tests to reach a overall conclusion, typically how to get a combined statistics from the statistics of individual hypothesis tests.
- How to coordinately perform *classification*, *hypothesis test*, and *feature selection*. E.g., can we have a same performance measure for both classification and hypothesis test? or otherwise how to trade off them satisfactorily.
- How to integrate multiple individual performances on either or both of classification and hypothesis test to get an overall good coordination between classification and hypothesis test.

Efforts on the first two aspects have also been partly discussed in [1] and studied also by many others [4-8]. The last two aspects are just preliminarily and incompletely involved in [1] and will be the focuses of this paper. In the next section, a preliminary A5 formulation in [1] is developed into a general formulation that integrates individual tasks of discriminative analysis by circularly implementing five basic actions. Then, it is shown in Sect.3 that this A5 formulation improves hypothesis testing with reliability enhanced (e.g., q-value reduced considerably).

Being different from conventional parametric and nonparametric statistics that firstly comes up an overall structure for each population and then detects significant difference between the summarized overall descriptions, we propose to firstly detect difference between paired samples and check whether these differences can be summarized into a significant one. Sect.4 proposes a sample pairing delta based nonparametric statistics and discusses its application to case control study and especially joint SNPs analyses. Finally, Sect.5 further proposes a boundary based statistics for coordinately implementing classification and hypothesis testing.

2 Integrative Hypothesis Testing and A5 Formulation

The A5 formulation comes from a general problem solving paradigm, refined from the mechanisms embedded in Hough Transform (HT), Randomized Hough Transform (RHT) [2,3] and Multi-sets-learning [16]. As illustrated in Fig.1, taking line detection within one image as an example, the HT detection is featured by a circular flow featured by five basic *mechanisms* or *actions*. First, it starts from picking one pixel from image, which is an instance of the action named *acquisition* (shortly denoted as A-1) for sampling evidence or data from the world in observation. Second, the HT

maps the pixel picked into a line in its parameter space $\theta=\{a,b\}$, which is an instance of the action *allocation & assumption* (A-2) that allocates information contained in the picked pixel, featured by a distributed allocation of evidence along a line that represents a set of candidate assumptions in the parameter space. Third, the HT quantizes a window of the parameter space into a lattice on which every cell is placed with an accumulator. We add one score to those accumulators located on the candidate assumptions provided by A-2, which is an example of the action *accumulation & amalgamation* (A-3) for integrating evidences about these candidate assumptions. Next, we inspect the scores of all the accumulators and detect those, that pass some threshold or become local maxima, as the final candidate conclusions on detected lines, which is an instance of the fourth action *apex-seeking & assessment* (A-4) that decides one or a set of final candidates with their corresponding scores either locating at peaks or becoming bigger than a threshold. Finally, the HT tests whether each of final candidates can be regarded as a detected line. In general, the job is named *affirmation* (A-5) that assesses whether each of candidate conclusions should be either discarded or identified as a final conclusion.

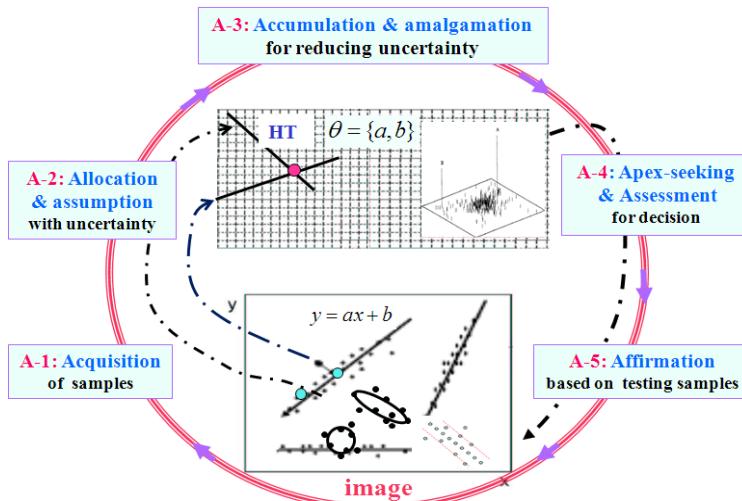


Fig. 1. A5 Formulation and One Exemplar

The tasks of object detection can be regarded as a special family of discriminant analysis that discriminates different populations of samples from geometrical shapes, while discriminant analysis usually refers to discriminate populations of samples from different statistical structures in term of either or both of *classification* and *hypothesis test*. To be specific, we consider a set of feature variables that represent the domain in observation, and then examine two populations within the domain via samples. Each sample is usually a vector or generally a structural set with each element being one specific value taken by the corresponding feature. Given one sample set $X^{(0)}$ from $P^{(0)}$ and the other sample set $X^{(1)}$ from $P^{(1)}$, we want to judge whether two populations are significantly different in term of *hypothesis test* and to classify samples into $P^{(0)}$ and $P^{(1)}$ with minimum confusion in term of *classification*. Usually, some features may be irrelevant, and some features are disturbed by noises or outliers. Instead of

considering $X^{(0)}$ and $X^{(1)}$ with all the features as a whole, both tasks are usually based on the task of finding an optimal subset of features in the name of *feature selection*.

Conventionally, the task pair of hypothesis test and feature selection is conducted separately from the task pair of classification and feature selection. In fact, two task-pairs are related but implemented under different measures that are actually not related monotonically. In sequel, we propose to use the A5 formulation to integrate multiple individual tasks of discriminative analyses with *classification of samples*, *test of hypotheses*, and *selection of features* performed coordinately.

Given two original sets $X^{(0)}$ and $X^{(1)}$ with each in an $M \times N$ matrix the consists of samples as rows. Let Ω to be a set of accumulation cells with each cell storing the evidence that supports the corresponding candidate assumption. Initially, Ω is empty, and gradually new cells are added in as new candidate assumptions come. Among Ω , a subset Ω_F is selected to store the final candidates. Moreover, we get $Z^{(0)}$ and $Z^{(1)}$ as a pair of testing sets on the domain of Ω_F . The samples of $Z^{(0)}$ and $Z^{(1)}$ maybe new ones from $P^{(0)}$ and $P^{(1)}$ or just randomly and partially picked from $X^{(0)}$ and $X^{(1)}$.

Schematically, the A5 formulation is featured by circularly implementing the following five *actions*.

Action A-1 (Acquisition of Evidence). Randomly picks m rows and n columns from $X^{(0)}$ and $X^{(1)}$ to form a pair of $m \times n$ matrices $Y^{(0)}$ and $Y^{(1)}$.

Remarks: Typically, the number of rows in $X^{(i)}$ is greatly larger than the number of columns in $X^{(i)}$ in real applications, which makes it unreliable to get a selection among a great number of rows merely based on a small number of columns. Randomly selecting $Y^{(i)}$ out of $X^{(i)}$ is a way of increasing an effective number of columns in a self-boosting manner, though suffering an over-optimistic risk.

Action A-2 (Allocation and Assumption). A subset of features $\{\omega\} = \{\omega_j, j=1, \dots, m\}$, corresponding to the m rows of $Y^{(i)}$, are selected by a SELECTOR with a score $s_{\{\omega\}}$, and then each ω^* is allocated a score s_{ω^*} according to its importance. Shown in Fig.2(a) is one example that is featured by a monotonic curve obtained by an ALLOCATOR.

Remarks: As an example of SELECTOR, we implement a sparse LDA learning [14] based on $Y^{(0)}$ and $Y^{(1)}$, resulting in a feature set $\{\omega\}$. We may even simply pick a set $\{\omega\}$ randomly. Usually, the allocated score $s_{\{\omega\}}$ is obtained in one of the following two ways:

Way A design an optimal classifier on the feature set $\{\omega\}$, and use its correct classification rate as the score $s_{\{\omega\}}$.

Way B make a multivariate hypothesis test (e.g., a Hotelling T² test [17]) based on $\{\omega\}$, and compute its corresponding q-value $q_{\{\omega\}}$ [4-6]. Then, we use $1 - q_{\{\omega\}}$ as the score $s_{\{\omega\}}$.

One rough treatment for ALLOCATOR is simply letting $s_{\omega} = s_{\{\omega\}} / \#\{\omega\}$ for each ω in $\{\omega\}$, where $\#A$ denotes the number of elements in the set A . A better choice is measuring the importance of a feature by its role in one sequential reduction. That is, each time a least important ω^* is removed with its score s_{ω^*} obtained as illustrated in Fig.2(a), with help of the following special treatments:

- if there are more than one feature ω^* with a same score s_{ω^*} , discard the one with the biggest p-value or misclassification rate obtained merely on this feature.
- if $s_{\omega^*} \leq 0$, we discard the least important pair of features by examining all the possible pairs in $\{\omega\}$, and so on so forth.

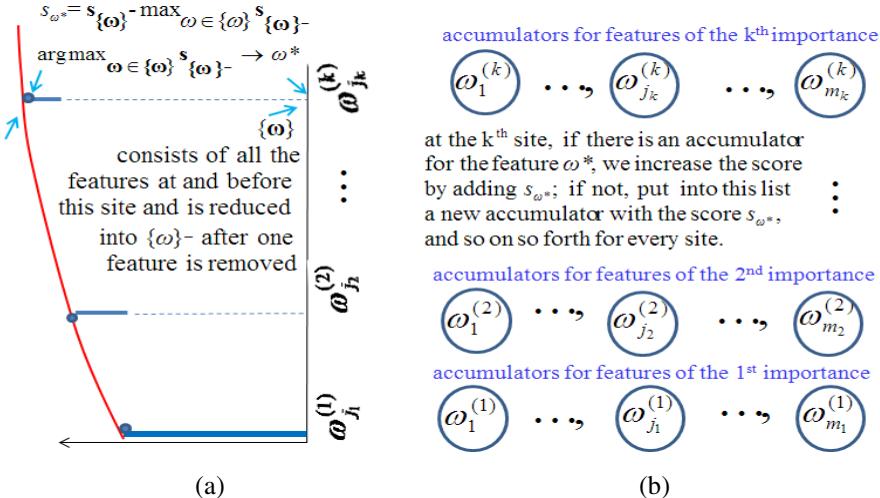


Fig. 2. Importance of features and accumulation of evidences (a) features are sequentially ordered by importance with the one of the 1st importance at the bottom and the curve indicates the accumulated score from the bottom up; (b) each of the ordered features is associated with a list of accumulators and each accumulator is associated with a different feature. Each list is empty at beginning and gradually added with accumulators as above described.

Action A-3 (Accumulation and Amalgamation). Evidences of importance are accumulated for the features $\{\omega\}$ provided by A-2, as shown in Fig.2(b).

Action A-4 (Apex-Seeking and Assessment). We pick a feature set $\Omega_F = \{\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(k)}\}$ in either of the following three ways:

- (1) Among the union of the 1st, 2nd, ..., kth rows with the scores for a same feature added together, we pick ones with the first k largest scores as Ω_F .
 - (2) Within the first row (at the bottom), pick the one associated with the largest score as $\omega^{(1)}$. Similarly, $\omega^{(2)}$ ($\neq \omega^{(1)}$) is picked from the top scored one within the second row, and so on so forth. Finally, $\omega^{(k)}$ is picked from the top scored one within the k^{th} row (excluding ones that duplicate the already picked features).
 - (3) After getting $\omega^{(1)}$ as above, $\omega^{(2)}$ ($\neq \omega^{(1)}$) is picked from the second row such that an optimal classifier based on $\omega^{(1)}, \omega^{(2)}$ results in the best classification rate as one feature associated with the largest score as $\omega^{(2)}$, and so on so forth.

Action A-5 (affirmation). Make hypothesis test or classification on $Z^{(0)}, Z^{(1)}$ based on Ω_F .

Remarks: Some applications verify the performance of classification, while other applications verify the performance of *hypothesis test*. There are also tasks that need to verify both types of performances too. Also, we may evaluate the importance of each $\omega \in \Omega_F$ in a way similar to that shown in Fig.2(a).

The above A5 formulation will increase the reliability of the final outcomes in Ω_F by integrating analyses made on the randomly picked sample sets at A-2, which shares a common point with the methods such as cross-validation, boosting, bagging, and stacking, which are widely studied in machine learning. In the next section, a probabilistic analysis will be provided to show that the evidence accumulation by A-3 and A-4 will significantly bring down the false discovery rate of hypothesis test [4-6].

The above A5 formulation seeks consensus on the importance and stability of features, while those machine learning studies seek consensus on the classification results instead of answering which features are important in their contributions to a good performance. Even worse, a combination of multiple classifiers with each classifier using different features actually enlarge the size of feature set considerably. This is unfavorable to those real tasks (especially in bioinformatics) that need to identify which features are mainly responsible to the final outcomes.

3 False Discovery Rate and Accumulated Reliability

Some probabilistic analyses are made on understanding how the A5 formulation can considerably improve the reliability. Starting from the basic concepts in Table 1, we summarize typical probability measures about hypothesis test from a Bayesian perspective in Table 2. On the last column, π_1 is the priori that H_0 is false, and π_0 is the priori that H_0 is true, while the last row lists the probabilities that the test rejects H_0 and fails to reject H_0 , respectively. The most widely used measure is the probability of Type I error, listed in the column of “reject H_0 ” on the same row of π_0 , which is described by the false positive probability usually called p -value that is controlled to be below an α level of significance.

Table 1. Basis concepts and notations in hypothesis test

	reject H_0	fail to reject H_0
Null H_0 is false	True positive	Type II error False negative
Null H_0 is true	Type I error False positive	True negative

(a)

	reject H_0	fail to reject	
H_0 false	<i>S</i>	<i>T</i>	d_1
H_0 true	<i>V</i>	<i>U</i>	d_0
	<i>R</i>	<i>W</i>	<i>d</i>

(b)

Instead of controlling the p value at the level α for each test, the familywise error rate is suggested to control the probability of making one or more Type I errors at the level α for all the hypotheses. However, this error rate is much too strict, especially when the number of hypotheses is large, and thus replaced by the false discovery rate (FDR) that is the expected proportion of false positive among all discoveries (rejected null hypotheses)[4]. As listed in a pair with the p -value in Tab.2, the q -value is actually the posteriori counterpart of the p -value from a Bayesian perspective [5,6]. With wide applications in big data analyses (e.g., genomics), FDR becomes a hot topic in the past one or two decades [4-6].

Table 2. Typical performance measures and Bayesian perspective

	Reject H_0 if ζ falls in its rejection region Γ	Fail to reject H_0 if $\zeta \notin \Gamma$	
H_1 $(H_0 \text{ is false})$	$E\left[\frac{S}{d}\right] = \pi_1 p(\zeta \in \Gamma H_1)$ $= p(H_1 \zeta \in \Gamma) p(\zeta \in \Gamma)$ Sensitivity or Power $\begin{cases} p(\zeta \in \Gamma H_1) = E\left[\frac{S}{d_1}\right] \geq 1 - \beta \\ p(H_1 \zeta \in \Gamma) = E\left[\frac{S}{R}\right] \end{cases}$	$E\left[\frac{T}{d}\right] = \pi_1 p(\zeta \notin \Gamma H_1)$ $= p(H_1 \zeta \notin \Gamma) p(\zeta \notin \Gamma)$ Type II error $\begin{cases} p(\zeta \notin \Gamma H_1) = E\left[\frac{T}{d_1}\right] \leq \beta \\ p(H_1 \zeta \notin \Gamma) = E\left[\frac{T}{W}\right] \end{cases}$	$\pi_1 = E\left[\frac{d_1}{d}\right]$
H_0 is true	$E\left[\frac{V}{d}\right] = \pi_0 p(\zeta \in \Gamma H_0)$ $= p(H_0 \zeta \in \Gamma) p(\zeta \in \Gamma)$ Type I error $\begin{cases} p = p(\zeta \in \Gamma H_0) = E\left[\frac{V}{d_0}\right] \leq \alpha \\ q = p(H_0 \zeta \in \Gamma) = E\left[\frac{V}{R}\right] \end{cases}$	$E\left[\frac{U}{d}\right] = \pi_0 p(\zeta \notin \Gamma H_0)$ $= p(H_0 \zeta \notin \Gamma) p(\zeta \notin \Gamma)$ Specificity $\begin{cases} p(\zeta \notin \Gamma H_0) = E\left[\frac{U}{d_0}\right] \geq 1 - \alpha \\ p(H_0 \zeta \notin \Gamma) = E\left[\frac{U}{W}\right] \end{cases}$	$\pi_0 = E\left[\frac{d_0}{d}\right]$
	$E\left[\frac{R}{d}\right] = p(\zeta \in \Gamma) = \pi_1 p(\zeta \in \Gamma H_1) + \pi_0 p(\zeta \in \Gamma H_0)$	$E\left[\frac{W}{d}\right] = p(\zeta \notin \Gamma) = 1 - p(\zeta \in \Gamma)$	

The Bayesian perspective can be extended to all the rest concepts listed in Tab.2, that is, we may have the posteriori counterparts of Type II error, sensitivity, and specificity. It follows from $p(H_1 | \zeta \in \Gamma) = 1 - p(H_0 | \zeta \in \Gamma)$ and $p(\zeta \in \Gamma | H_1) = 1 - p(\zeta \notin \Gamma | H_1)$ that these measures are related, which provides not only insights but also alternative way to compute the q-value. When we get some structure about H_1 , it is more feasible to estimate $p(\zeta \notin \Gamma | H_1)$ and thus its posteriori counterparts. Also, the q-value may be estimated from the likelihood ratio positive (LR+), which has been studied in diagnostic testing of evidence-based medicine. It follows from Tab.2 that we have

$$q = p(H_0 | \zeta \in \Gamma) = \frac{\pi_0 p(\zeta \in \Gamma | H_0)}{\pi_1 p(\zeta \in \Gamma | H_1) + \pi_0 p(\zeta \in \Gamma | H_0)} = \frac{1}{LR_+ \pi_1 / \pi_0 + 1}, \quad (1)$$

$$LR_+ = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{p(\zeta \in \Gamma | H_1)}{p(\zeta \in \Gamma | H_0)} \quad \text{or} \quad LR_+ = \frac{\pi_0}{\pi_1} \frac{1}{q} - 1,$$

from which we observe that a bigger LR_+ value actually means a smaller q value.

Next, we explain that the A5 formulation brings down the q-value significantly by A-3 and A-4, and enhances the performances of hypothesis test. For simplicity, we assume that each time every accumulator adds by a score 1 at A-3 and that the test per circling is independent from each other. Though this assumption is actually not true

because samples are randomly picked from the same $X^{(0)}, X^{(1)}$ at A-1, we may get some insights by this rough approximation.

We observe a particular cell $\omega_f \in \Omega$ that just reached a threshold τ after M times of circling during the A5 implementation. Precisely, the statistics in consideration is the set $\{\omega_f\}$ of all the cells in Ω with its region of rejection as follows:

$$\Gamma^\tau = \bigcup_{\omega_f \in \Omega} \Gamma_{\omega_f}^\tau, \quad \Gamma_{\omega_f}^\tau = [\tau, \tau+1, \dots, +\infty), \quad (2)$$

which means that there is at least one $\omega_f \in \Gamma_{\omega_f}^\tau$.

It follows from the last row of Tab.2 that the probability for $\{\omega_f\}$ to fall in its rejection region is given as follows:

$$P(\{\omega_f\} \in \Gamma^\tau) = C_M^\tau p^\tau (\zeta \in \Gamma) [1 - p(\zeta \in \Gamma)]^{M-\tau}, \quad (3)$$

and its corresponding P-value is correspondingly given as follows:

$$P = P(\{\omega_f\} \in \Gamma^\tau | H_0) = C_M^\tau p^\tau (1-p)^{M-\tau} < C_M^\tau p^\tau, \quad p = p(X \in \Gamma | H_0). \quad (4)$$

In other words, the score accumulation at A-3 can bring down the P-value when $C_M^\tau p^{\tau-1} < 1$, which is possible as long as the threshold τ becomes large enough.

Moreover, it is promising to observe the Q-value by the Bayes posteriori:

$$Q = P(H_0 | \{\omega_f\} \in \Gamma^\tau) = \frac{P(H_0) P(\{\omega_f\} \in \Gamma^\tau | H_0)}{P(\{\omega_f\} \in \Gamma^\tau)} = \frac{\pi_0^M C_M^\tau p^\tau (1-p)^{M-\tau}}{C_M^\tau p^\tau (\zeta \in \Gamma) [1 - p(\zeta \in \Gamma)]^{M-\tau}} \quad (5)$$

$$= q^\tau p(H_0 | X \notin \Gamma)^{M-\tau} < q^\tau, \quad \text{with } P(H_0) = \pi_0^M \text{ and } q = p(H_0 | X \in \Gamma).$$

That is, score accumulation indeed brings down the Q-value simply as long as $\tau > 1$.

Such accumulated reliability increases as the threshold τ becomes larger. However, there is no free lunch. It follows from Eq.(3) that the probability of rejecting H_0 also reduces considerably, which leads to either the risk of missing significant features (i.e., detecting power reduced) for a fixed number M or a large computing cost for keeping A5 circling until finding enough significant features.

4 Sample Pairing Delta and Case Control Study

Hypothesis test focuses at evaluating a deviation from the hull assumption:

$$H_0 : \text{there is no difference between } P^{(0)} \text{ and } P^{(1)}. \quad (6)$$

Taking multivariate Gaussian population as an example, we consider

$$X^{(\ell)} = [x_1^{(\ell)}, \dots, x_{N_\ell}^{(\ell)}], \quad \ell = 0, 1, \quad x_t^{(\ell)} \text{ from } G(x | \alpha^{(\ell)}, \Sigma). \quad (7)$$

Under the null hypothesis, we get the following maximum likelihood estimation:

$$\begin{aligned} \alpha^{(0)} \alpha^{(0)} + \alpha^{(1)} \alpha^{(1)}, \quad \Sigma = \alpha^{(0)} \Sigma^{(0)} + \alpha^{(1)} \Sigma^{(1)}, \quad \alpha^{(\ell)} = N_\ell / N, \quad N = \sum_\ell N_\ell, \\ \alpha^{(\ell)} = \frac{1}{N_\ell} \sum_{t=1}^{N_\ell} x_t^{(\ell)}, \quad \Sigma^{(\ell)} = \frac{1}{N_\ell - 1} \sum_{t=1}^{N_\ell} (x_t - \alpha^{(\ell)}) (x_t - \alpha^{(\ell)})^T. \end{aligned} \quad (8)$$

Typically, the deviation from H_0 is measured by the Hotelling T^2 statistics [17]:

$$\begin{aligned} T^2(\theta_0, \theta_1) &= \alpha_0 \alpha_1 (\boldsymbol{\alpha}^{(0)} - \boldsymbol{\alpha}^{(1)})^T \Sigma^{-1} (\boldsymbol{\alpha}^{(0)} - \boldsymbol{\alpha}^{(1)}) \\ &= \frac{\alpha_0 \alpha_1}{\alpha_0^2 + \alpha_1^2} [(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha})^T \Sigma^{-1} (\boldsymbol{\alpha}^{(1)} - \boldsymbol{\alpha}) + (\boldsymbol{\alpha}^{(0)} - \boldsymbol{\alpha})^T \Sigma^{-1} (\boldsymbol{\alpha}^{(0)} - \boldsymbol{\alpha})], \quad \theta_i = \{\alpha_i, \boldsymbol{\alpha}^{(i)}, \Sigma^{(i)}\}, \end{aligned} \quad (9)$$

which measures the deviation of $P^{(0)}, P^{(1)}$ from $G(x|\mu, \Sigma)$ as a reference about H_0 .

Other than parametric statistics, the Kolmogorov-Smirnov test and the Mann-Whitney U test are two general nonparametric methods to test whether samples come from the same distribution [15]. The former uses the maximal distance between cumulative frequency distributions of these two samples as the statistics, while the latter takes the difference between mean ranks of these two samples as the statistics.

Here, we propose another nonparametric statistics as follows:

$$\begin{aligned} \zeta_A &= D(X^{(0)} \| X^{(1)}), \\ D(A \| B) &= \frac{1}{\#A \#B} \sum_{x \in A} \sum_{y \in B} \delta^T(x, y) \delta(x, y), \quad x = [a_1, \dots, a_d], \quad y = [b_1, \dots, b_d], \end{aligned} \quad (10)$$

where $\delta(x, y)$ indicates the variation by pairing the sample x with y , and is called sample-pairing-delta. One special case considers element by element independently:

$$\delta(x, y) = [\delta_1(a_1, b_1), \dots, \delta_d(a_d, b_d)]^T,$$

where $\delta_j(a_j, b_j)$ measures the variation from a_j to b_j . One example is that elements are homogenous up to unknown scales w_j 's, that is, we have

$$\delta_j(a_j, b_j) = w_j \delta(a_j, b_j). \quad (11)$$

In general, each element of $\delta(x, y)$ depends on vector x and vector y . Typically, we may consider that each element comes from $\delta(a_j, b_j), j=1, \dots, d$ by a linear map

$$\delta(x, y) = W[\delta(a_1, b_1), \dots, \delta(a_d, b_d)]^T, \quad (12)$$

where W is a transformation matrix. Particularly, when $\delta(x, y) = x - y$ and $W = \Sigma^{-0.5}$, we get the following Mahalanobis distance:

$$\delta^T(x, y) \delta(x, y) = (x - y)^T \Sigma^{-1} (x - y). \quad (13)$$

This sample-pairing-delta $\delta(x, y)$ based statistics ζ_A provides a practical facility too. In real applications, there are frequently samples that have elements with missing values. Traditionally, we either discard such a sample or fill the missing values by certain estimation of the missing value based on other samples. On one hand, discarding the sample makes the small sample size problem become more serious, as encountered in bioinformatics. On the other hand, it is difficult to estimate a missing value well and usually a rough estimation leads to a bad performance.

We believe that a best policy is letting the affect of missing value to other parts of computation to be as less as possible. E.g., in the SNP analysis [13], $\delta_j(a_j, b_j)$ measures the extent of a variation from a_j into b_j . Typically, a SNP variation occurs rarely and thus if it is missed in detection we can conservatively regard that there is no variation. That is, we may simply consider a rectified scale in Eq.(11) by

$$w_j = \begin{cases} 0, & \text{at least one of } a_j, b_j \text{ is missing,} \\ 0, & \text{when } a_j = b_j, \\ \neq 0, & \text{otherwise.} \end{cases} \quad (14)$$

The simplest case is $w_j=1$, or the degenerated case $W=I$ in Eq.(12). In general, W in Eq.(12) may consider the dependence cross elements by

$$W = C_v^{-0.5}, \quad C_v = C_v(X^{(0)} \cup X^{(1)} \parallel X^{(0)} \cup X^{(1)}), \quad (15)$$

$$C_v(A \parallel B) = \frac{1}{\#A \# B} \sum_{x \in A} \sum_{y \in B} [\delta_1(a_1, b_1), \dots, \delta_d(a_d, b_d)] [\delta_1(a_1, b_1), \dots, \delta_d(a_d, b_d)]^T \text{ at } W=I,$$

where the matrix C_v measures co-variations across elements. Assuming independence cross elements, W could be given as follows

$$W = \text{diag}[w_1, \dots, w_d] = \text{diag}[C_v^{-0.5}], \text{ or even simply } W = (d / \sqrt{\text{Trace}[C_v]}) I. \quad (16)$$

In implementation, we usually do not know the distribution of ζ_A and thus cannot estimate the p-value in a standard way. Still, we may approximately estimate the p-value by a Monte Carlo simulation under the null H_0 by Eq.(6), e.g., we make the following permutation test:

$$P_{\text{value}} = \frac{1}{\#\Pi} \{1 + \sum_{\pi \in \Pi} I[D(X_\pi^{(0)} \parallel X_\pi^{(1)}) > D(X^{(0)} \parallel X^{(1)})]\}, \quad (17)$$

where $I(a>b)=1$ if $a>b$, otherwise $I(a>b)=0$, and Π is the set of all the possible permutations, with each $\pi \in \Pi$ turning out $X_\pi^{(0)}, X_\pi^{(1)}$ [11].

Next, we proceed to address other favorable features of the statistics ζ_A by Eq.(10) in comparison with typical studies of hypothesis test.

First, in a standard way, a statistics and its distribution are derived under the null H_0 by Eq.(6), after which the value of this statistics is computed from samples and a p-value is estimated according to the distribution to test whether this H_0 breaks significantly. However, it is challenging to estimate the distribution of this statistics and to compute the p-value according to this distribution, which makes the standard approaches of hypothesis test limited to only a few commonly assumed distributions. In contrast, the statistics ζ_A by Eq.(10) makes hypothesis test performed in an alternative way that directly measures the difference between $P^{(0)}$ and $P^{(1)}$ regardless the null H_0 , and compute the p-value by a Monte Carlo simulation under the null H_0 without estimating the distribution of ζ_A .

A parametric statistics may also be obtained in a similar way. Firstly in [10] and subsequently in [1], the following Kullback–Leibler (KL) divergence is suggested

$$\zeta_{KL} = KL(p(X \mid \Theta^{(1)}) \parallel p(X \mid \Theta^{(0)})), \quad KL(p \parallel q) = \int p(x) \ln[p(x)/q(x)] dx, \quad (18)$$

as a general formulation of statistics that aims at the difference between $P^{(0)}$ and $P^{(1)}$, where $\Theta^{(1)}$ is an estimation of Θ based on $X^{(1)}$ while $\Theta^{(0)}$ is an estimation of Θ on $X^{(0)}$. By the way, getting $\Theta^{(0)}$ estimated on both $X^{(0)}$ and $X^{(1)}$, ζ_{KL} also leads to statistics in a standard sense. E.g., let q, p given by $G(x \mid \mu^{(1)}, \Sigma)$, $G(x \mid \mu^{(0)}, \Sigma)$ in Eq.(18), we are lead to the Hotelling statistics by Eq.(9), with details referred to [1] (especially p870).

Second, typical existing statistics, derived either parametrically (e.g., the Hotelling T^2 statistics [17] and those frequency table based tests used for SNP analyses) or non-parametrically (e.g., the Kolmogorov-Smirnov test and the U test), shares a common feature. That is, the overall structure or description of $P^{(i)}$ is firstly summarized from its own samples, respectively for $i=0,1$, and then, the summarized overall structure of $P^{(1)}$ is compared with the summarized overall structure of $P^{(0)}$ to detect whether there is a significant difference. In contrast, the statistics ζ_A by Eq.(10) starts to detect delta $\delta(x,y)$ (i.e., difference) between paired samples x,y and then summarize these deltas for detecting a significant difference between $X^{(0)}$ and $X^{(1)}$.

Such a sample-pairing-delta based statistics ζ_A may have some asymptotic relation to a standard *hypothesis test* method. For the example by Eq.(13), it can be analytically shown that as the sample size $N^{(0)}+N^{(1)}$ becomes large enough we have

$$\zeta_A - d \rightarrow T^2(\theta_0, \theta_1) \text{ as } N^{(1)} + N^{(0)} \rightarrow \infty, \quad (19)$$

where $T^2(\theta_0, \theta_1)$ is the Hotelling statistics by Eq.(9), i.e., the difference between means prevails. In other words, ζ_A by Eq.(10) differs from the standard hypothesis test especially for detecting population difference on a finite size of samples.

Third, typical statistics such as the Hotelling T^2 statistics and those frequency table based statistics actually test a deviation from the null H_0 by Eq.(6) via testing the differences between mean values subject to certain constraints. That is, the statistics becomes zero when there is no difference between mean values. In contrast, ζ_A by Eq.(10) will be still a nonzero value $\zeta_0 \neq 0$ in such a case of no difference. For some distributions, e.g., $G(x|\mu^{(1)}, \Sigma) = G(x|\mu^{(0)}, \Sigma)$, this $\zeta_0 \neq 0$ tends to asymptotically a constant as shown in Eq.(19) and thus does not contain any useful information about difference between $P^{(0)}$ and $P^{(1)}$. Even so, it has no bad effect on the permutation test by Eq.(17) because inequality will not be affected by adding a constant in both the sides. For two populations $P^{(0)}$ and $P^{(1)}$ with no difference between mean values but still some higher order difference, ζ_A by Eq.(10) may detect some information about this difference. In other word, this ζ_A is more powerful than those mean-value based statistics for detecting a deviation from the null H_0 by Eq.(6).

The last but not least, we consider the case-control studies that are widely encountered in real applications of hypothesis test. The samples $X^{(0)}$ of $P^{(0)}$ come from a normal population as benchmark or called control. There is usually a reference point about the normal by which we calibrate each feature to represent its deviation from the corresponding reference. For control samples, deviations from the reference are usually isotropic random noises with zero mean. On the other hand, samples $X^{(1)}$ of $P^{(1)}$ come from one abnormal population or called case, with a systematic deviation from the reference. The task of case-control studies aims at detecting this systematic deviation. Typical statistics, especially those based on the difference between mean values, consider $P^{(0)}$ and $P^{(1)}$ in a same distribution form $p(X|\Theta)$ but different in the values that Θ takes. With help of Eq.(18), we may further proceed to consider $P^{(0)}$ and $P^{(1)}$ in different distribution forms.

Moreover, extensions may also be made to consider that the reference may not necessarily represent the normal and that deviations of control samples from the reference may not necessarily isotropic noises. Taking GWAS analysis [13] as an example, a case sample takes a symbol η and a control sample takes a symbol ξ at one SNP site. At this site, both η and ξ may take the value 0 for SS as a reference, the value 1 for the variation Ss , and the value 2 for the variation ss .

Generally speaking, deviations from the reference can be classified into two types, one by control samples and one by case samples. It is insightful to watch which type dominates, for which we further look into the following two scenarios:

(1) $P(\eta > \xi)$, i.e., the case samples deviate more badly than control samples do. E.g., for the SNP analysis, a disease is likely caused by certain variations. We may measure the difference between $X^{(0)}$ and $X^{(1)}$ in this scenario by

$$\begin{aligned}\zeta_A^{1>0} &= \sum_j w_j \beta_j^{(0)} \beta_j^{(1)} D_j^{1>0}, \quad D_j^{1>0} = \frac{1}{N_j^{(0)} N_j^{(1)}} \sum_{\xi \in X_j^{(0)}} \sum_{\eta \in X_j^{(1)}} P(\eta > \xi) \delta^2(\xi, \eta), \\ \beta_j^{(i)} &= N_j^{(i)} / \sum_j N_j^{(i)}, \quad w_j = 1/D_{X_j^{(0)} \cup X_j^{(1)}}, \quad D_A = \frac{0.5}{\# A \# A} \sum_{\xi \in A} \sum_{\eta \in A} P(\xi \neq \eta) \delta^2(\xi, \eta), \\ N_j^{(i)} &= \# X_j^{(i)} \text{ (excluding missing elements)}, \quad X_j^{(i)} \text{ from the } j\text{th column of } X^{(i)}.\end{aligned}\quad (20)$$

The vectors x, y in Eq.(10) correspond to one pair of samples from one computational unit (e.g., a gene), with the j th element pair ξ and η of x, y representing the j th site.

(2) $P(\eta < \xi)$, i.e., the control samples deviate more badly than case samples. E.g., the normal population experienced variations to adapt an environmental change, while a disease is likely caused due to a lack of such variations. We may similarly get $\zeta_A^{0>1}$ simply with $\eta > \xi$ in Eq.(20) replaced by $\eta < \xi$ while $1 > 0$ in Eq.(20) replaced by $0 < 1$.

Moreover, we may also get

$$\begin{aligned}D(X^{(0)} \| X^{(1)}) &= \zeta_A^{1>0} + \zeta_A^{0>1} = \sum_j w_j \beta_j^{(0)} \beta_j^{(1)} D_j, \\ D_j &= \frac{1}{N_j^{(0)} N_j^{(1)}} \sum_{\xi \in X_j^{(0)}} \sum_{\eta \in X_j^{(1)}} P(\eta \neq \xi) \delta^2(\xi, \eta).\end{aligned}\quad (21)$$

Finally, we may put each of the above $\zeta_A^{0>0}$, $\zeta_A^{0>1}$ and $D(X^{(0)} \| X^{(1)})$ into Eq.(17) to implement permutation test.

5 Boundary Based Statistics

The task of discriminant analysis is selecting a subset of features on which we observe either a best separation between $P^{(0)}$ and $P^{(1)}$ in term of *classification* or a significant overall difference between $P^{(0)}$ and $P^{(1)}$ in term of *hypothesis test*.

For a best separation, we consider a boundary that separates samples from $P^{(0)}$ and $P^{(1)}$ with a least number of misclassified samples, where a misclassified sample is one that comes from one population but be classified into the other. Whether two populations could be well separated relates to whether there is a significant overall difference between the populations. However, two concepts are not same. Also, classification and hypothesis test are implemented under different performance measures that are not monotonically related. The one for classification focuses on boundary separation, while the one for hypothesis test focuses on difference of the overall structures. A best performance for one may not be necessarily the best for the

other. Conventionally, classification and hypothesis test are studied separately. Here, we attempt to integrate the two subtasks by reexamining the role of discriminant boundary in getting statistics for hypothesis test.

Observing Eq.(8), μ actually acts as a boundary that separates two populations, and T^2 measures the Mahalanobis distance to this boundary from both the sides of $\mu^{(0)}$ and $\mu^{(1)}$. A boundary is also implied between x and y in Eq.(13). However, such a rough boundary only results in a rough separation of $P^{(0)}$ and $P^{(1)}$. Thus, we are motivated to embed an optimal best boundary into one statistics in order to judge whether two populations are significantly different.

One rather straightforward way is a two step implementation as follows.

Step 1: use a Bayes classifier or one alternative to separate samples into

$$X^{(0)} = X_1^{(0)} \cup X_0^{(0)}, \quad X^{(1)} = X_0^{(1)} \cup X_1^{(1)}, \text{ with a confusion matrix } \begin{bmatrix} n_{00}, & n_{01} \\ n_{10}, & n_{11} \end{bmatrix}, \quad (22)$$

where $X_i^{(j)}$ consists of samples that come from $X^{(j)}$ and classified into the population $P^{(i)}$, with $n_{ji} = \#X_i^{(j)}$.

Step 2: measure the deviation from the null hypothesis H_0 by

$$\zeta_B = \frac{n_{00}n_{11}D(X_0^{(0)} \parallel X_1^{(1)})}{n_{00}n_{01}D(X_0^{(0)} \parallel X_1^{(0)}) + n_{10}n_{11}D(X_0^{(1)} \parallel X_1^{(1)})}, \quad (23)$$

where $D(A||B)$ is same as defined in Eq.(10). Alternatively, we may replace this nonparametric statistics by a parametric statistics. Considering the asymptotic approximation by Eq.(19), we may let that

$$D(X_0^{(j)} \parallel X_1^{(\ell)}) \text{ is replaced with } T^2(\theta_0^{(j)}, \theta_1^{(\ell)}) + d, \quad (24)$$

where $\theta_i^{(j)}$ is a maximum likelihood estimation of a Gaussian distribution based on the subset $X_i^{(j)}$ of samples. It follows from Eqn.(5) in [1] that we may also use

$$T^2(\theta_0^{(j)}, \theta_1^{(\ell)}) = \frac{n_{j0}n_{\ell 1}}{(n_{j0} + n_{\ell 1})^2} KL(X_0^{(j)} \parallel X_1^{(\ell)}). \quad (25)$$

Still, the above two step implementation is made under two different measures, i.e., misclassification at Step 1 and ζ_B at Step 2.

Instead, we can also use a same measure to implement two steps coordinately, for which we estimate θ to maximize a parametric $\zeta_B(\theta)$ that varies with a discriminant boundary described by a parameter θ .

Given a boundary $f(x, \theta)=0$ such that a sample x is classified into $P^{(1)}$ if $f(x, \theta)>0$ and into $P^{(0)}$ if $f(x, \theta)<0$, one example of such a parametric $\zeta_B(\theta)$ is given as follows:

$$\zeta_B(\theta) = D_{separa}(\theta) / D_{confus}(\theta), \quad (26)$$

$$D_{confus}(\theta) = \sum_{y \in X^{(1)}, f(y, \theta) < 0} s(d_f(y, \theta)) + \sum_{x \in X^{(0)}, f(x, \theta) > 0} s(d_f(x, \theta)),$$

$$D_{separa}(\theta) = \sum_{y \in X^{(1)}, f(y, \theta) > 0} s(d_f(y, \theta)) + \sum_{x \in X^{(0)}, f(x, \theta) < 0} s(d_f(x, \theta)),$$

where $s(r)$ is a monotonically increasing function with respect to a scalar variable r , and $d_f(x, \theta)$ denotes the shortest distance of a sample x to $f(x, \theta)=0$ by

$$d_f(x, \theta) = \sqrt{\min_{u, f(u, \theta)=0} \|x - u\|^2}. \quad (27)$$

When $s(r)=0$ for $r=0$ otherwise $s(r)=1$ for $r>0$, it follows from Eq.(26) that

$$\zeta_B(\theta) = (n_{11} + n_{00}) / (n_{10} + n_{01}) = -1 + (n_{10} + n_{01})^{-1}, \quad (28)$$

which is monotonically decreasing with respect to the misclassification error, and its maximization is equivalent to minimizing the misclassification. On the other hand, when $s(r)=r^2$, $\zeta_B(\theta)$ by Eq.(26) is conceptually a counterpart of ζ_A by Eq.(10), Eq.(15), and Eq.(16). In other words, $\zeta_B(\theta)$ by Eq.(26) covers *classification* and *hypothesis test* as two special cases, and generally trades off the natures of both.

We may emphasize one over other by different choices of $s(r)$, e.g., we consider

$$s(r) = r^\alpha, \quad \alpha > 0, \quad (29)$$

for which we are leaded to hypothesis test as $\alpha=2$ and to classification as $\alpha\rightarrow 0$, as well as to one intermediate case when α takes a value between (0,2].

In implementation, one bottleneck is solving Eq.(27) efficiently, though we simply have

$$d_f(x, \theta) = (x - c)^T w / \|w\|, \text{ for a linear boundary } f(u, \theta) = w^T(u - c) = 0. \quad (30)$$

Indirectly, $d_f(x, \theta)$ is able to be solved by the Lagrangian for a quadratic function $f(x, \theta)=0$ [12], for which one example is a Bayes classifier with the boundary below

$$f(x, \theta) = \ln[\alpha_1 G(x | \alpha^{(1)}, \Sigma^{(1)})] - \ln[\alpha_0 G(x | \alpha^{(0)}, \Sigma^{(0)})] = 0,$$

which is a quadratic equation of x and degenerates to a linear equation when $\Sigma^{(0)} = \Sigma^{(1)}$.

Instead of considering $f(x, \theta)=0$ as a linear or quadratic equation globally in the sample space, we may model a discriminant boundary by a mixture of linear functions $f(x, \theta_j)=0$, $j=1, 2, \dots, m$ as follows:

$$\begin{aligned} \zeta_B(\theta_j) &= \sum_j \zeta_B(\theta_j), \quad \zeta_B(\theta_j) = D_{separa}(\theta_j) / D_{confus}(\theta_j), \\ D_{confus}(\theta_j) &= \sum_{y \in X_k^{(1)}, f(y, \theta_j) < 0} s(d_f(y, \theta_j)) + \sum_{x \in X_k^{(0)}, f(x, \theta_j) > 0} s(d_f(x, \theta_j)), \\ D_{separa}(\theta_j) &= \sum_{y \in X_k^{(1)}, f(y, \theta_j) > 0} s(d_f(y, \theta_j)) + \sum_{x \in X_k^{(0)}, f(x, \theta_j) < 0} s(d_f(x, \theta_j)). \end{aligned} \quad (31)$$

That is, for each $f(x, \theta_i)=0$ we concentrate on considering the k nearest samples of $X^{(0)}$, $X^{(1)}$ to $f(x, \theta_j)=0$ as follows:

$$X_k^{(i)}(\theta_j) = \{k \text{ samples of } X^{(i)} \text{ with the minimum distances to } f(x, \theta_j) = 0\}, \quad i = 0, 1.$$

The overall separation of $X^{(0)}$, $X^{(1)}$ is described by combining m local liner discriminant boundaries. For a small k , the focus is on a border based misclassification with some ignorance of overall difference between populations. In other words, we get a better classification but a weak detecting power for hypothesis test. As k increases, the focus gradually switches to increasing the detecting power on the difference in overall structure, but suffering some classification accuracy. The value of k trades off between classification and hypothesis test. Also, we may control $s(r)$ by Eq.(29) for a similar role. E.g., $\alpha<0$ discounts samples away from the border.

For a classification task, samples around the border take major role on estimating the boundary structure between populations, while samples away from the border take a role of regularizing the estimation especially when there are few boundary samples.

How to decide an appropriate k in Eq.(31) remains a challenge. We start from a small k to iteratively maximize $\zeta_B(\theta)$ for learning θ , with k gradually increasing. Then, with the resulted θ , we test hypotheses on $X^{(0)}, X^{(1)}$ as k becomes large enough.

For a hypothesis testing task, we start from a large k and gradually reduce to a small k for an improved classification performance.

The statistics by Eq.(26) is different from ones by Eq.(9) and Eq.(10) that are made directly in the apparent domain (*shortly A-domain*), where samples are directly observed and overall structures are compared. Also, this statistics differs from ones that is featured by statistics computed indirectly in an inner domain (*shortly I-domain*) where samples are mapped into and where misclassification or separation is measured [1,10]. Instead, ζ_B by Eq.(26) integrates both the statistics $D(\|\cdot\|)$ in the *A-domain* and the statistics n_{ji} in an inner decision domain to measure the differences between two populations.

Last but not least, we need further theoretical understanding on the third issue in Sect.1. Namely, we want to know whether *best classification* and *best hypothesis test* will become asymptotically equivalent as the sample size tends to infinite, subject to a moderate condition or can be achieved under a same performance measure, e.g., under ζ_B by Eq.(31) with a same value of k ?

Acknowledgment. This work was supported by a CUHK Direct grant for 2013-2014.

References

1. Xu, L.: Matrix-variate discriminative analysis, integrative hypothesis testing, and geno-pheno A5 analyzer. In: Yang, J., Fang, F., Sun, C. (eds.) IScIDE 2012. LNCS, vol. 7751, pp. 866–875. Springer, Heidelberg (2013)
2. Xu, L.: A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition 40, 2129–2153 (2007)
3. Xu, L., Oja, E.: Randomized Hough transform. In: Encyclopedia of Artificial Intelligence, pp. 1354–1361. IGI Global, Hershey (2008)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. of Royal Statistical Society B 57(1), 289–300 (1995)
5. Storey, J.D.: A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B 64(3), 479–498 (2002)
6. Storey, J.D., Tibshirani, R.: Statistical significance for genome-wide studies. Proc. of the National Academy of Sciences 100(16), 9440–9445 (2003)
7. Glezko, G.V., Emmert-Streib, F.: Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics 25, 2348–2354 (2009)
8. Alves, G., Yu, Y.: Combining independent, weighted p-values: achieving computational stability by a systematic expansion with controllable accuracy. PLoS ONE 6(8), e22647 (2011)
9. Zaykin, D.V.: Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J. Evol. Biol. 24(8), 1836–1841 (2011)

10. Xu, L.: Semi-Blind Bilinear Matrix System, BYY Harmony Learning, and Gene Analysis Applications. In: Proc. of 6th International Conf. on New Trends in Information Science, Service Science and Data Mining, Taipei, October 23–25, pp. 661–666 (2012)
11. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer (2005)
12. Liu, Z.Y., Qiao, H., Xu, L.: Multisets mixture learning based ellipse detection. Pattern Recognition, 39731–39735 (2006)
13. Bansal, V., Libiger, O., Torkamani, A., Schork, N.J.: Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11, 773–785 (2010)
14. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics 53, 406–413 (2011)
15. Bagdonavicius, V., Kruopis, J., Nikulin, M.S.: Non-parametric tests for complete data. ISTE & WILEY, London & Hoboken (2011)
16. Xu, L.: Data smoothing regularization, multi-sets-learning, and problem solving strategies. Neural Networks 16, 817–825 (2003/2012)
17. Hotelling, H.: The generalization of Student’s ratio. Annals of Mathematical Statistics 2(3), 360–378 (1931)

Erratum: Camera Localization and Pose Estimation Using an RGBD Sensor

Hao Chen¹, Yan Yuan¹, John McDonald², and Thomas Whelan²

¹ Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education,
Beihang University, Beijing, China
jackiechensuper@gmail.com

² Department of Computer Science, National University of Ireland, Maynooth, Ireland

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 805–812, 2013.
© Springer-Verlag Berlin Heidelberg 2013

DOI 10.1007/978-3-642-42057-3_113

The paper starting on page 805 of this volume originally listed John McDonald and Thomas Whelan as authors. Their names have been removed, because they were not authors of this paper and their names were included without their permission. There may be older versions of this paper in circulation that include their names.

Author Index

- Abudureyimu, Halidan 40
Bai, Zhen Long 626
Balla-Arabé, Souleymane 529
Bian, Hui 6
Cai, Deng 238
Chang, Daqing 682
Chen, Changhong 192
Chen, Fanglin 128, 216
Chen, Fu-Hua 561
Chen, Hao 805
Chen, Heng 813
Chen, Huafu 554, 813
Chen, Wei 176, 230
Chen, Xian Zhong 70, 626
Chen, Xiaojun 481
Chen, Ying 666
Chen, Youbin 489, 521
Chen, Yu 223
Chen, Zenghai 62
Chen, Zhihui 441
Cheng, H.D. 538
Chi, Jiannan 880
Cui, Zhenchao 328
Dai, Xiubin 168
Deng, Hong 789
Deng, Renren 40
Deng, Yao 521
Deng, Yong 410
Dong, Jian 875
Du, Bo 336, 351
Duan, Xiping 427
Fan, Dandan 827
Fan, Jiulun 505
Fan, Zizhu 457
Fang, Tian-Zhu 513
Feng, Bingtao 410
Feng, Hui 771
Feng, Jufu 465
Feng, Yajuan 376
Feng, Yu-zhi 579
Fu, Dongmei 255, 263
Fu, Xin 70
Gan, Zongliang 192
Gao, Fei 184
Gao, Guanyin 160
Gao, Jun 435
Gao, Ming 649
Gao, Xianjun 570
Gao, Xinbo 184, 529, 697, 748
Gao, Yongsheng 844
Gao, Zi-Ming 705
Geng, Xin 603
Gong, Caixia 481
Gui, Jie 820
Guo, Jingyuan 771
Guo, Longyuan 1
Guo, Qi 393
Guo, Zhenhua 521
Han, Bing 748
Hao, Shuai 449
Hao, Zhanlong 489
He, Wu 160
He, Xiaofei 238
Hou, Qing Wen 70, 626
Hu, Bo 771
Hu, Dewen 128, 216, 295, 368, 410
Hu, Qiaoli 797
Hu, Ruijuan 120
Hu, Weiming 120, 136, 152, 449
Hu, Weisong 546
Hu, Yao 238
Huang, Bo-Shiang 14
Huang, Dong 112
Huang, Jianhua 538
Huang, Jianyu 649
Huang, Rong 344
Huang, Wenhao 546
Huang, Xin 351
Huang, Yongzhen 418, 603
Huang, Zhenhua 722
Jia, Yu 860
Jia, Yunde 30, 54

- Jiang, Jun 295, 368
 Jin, Hui 200
 Kang, Rui 835
 Kumar, Anand 22
 Lai, Jian-Huang 86, 112
 Lao, Wenchao 62
 Lei, Ying-Ke 820
 Li, Bing 120, 136
 Li, Congli 208
 Li, Deyuan 672
 Li, Feifei 611
 Li, Guoqiang 595
 Li, Jie 570, 697
 Li, Jun 152
 Li, Lei 401
 Li, Man 546
 Li, Ming 740
 Li, Qian 279
 Li, Qianwen 779
 Li, Sha 6
 Li, Tao 740
 Li, Weibing 62
 Li, Weifeng 521
 Li, Wuchen 263
 Li, Xiangang 473
 Li, Xiaogang 255
 Li, Xingen 263
 Li, Xuelong 697
 Li, Ying 587, 595, 860
 Li, Yirui 246, 852
 Li, Zong 705
 Liang, Zhizhen 176, 230
 Liao, Fangshun 587
 Liao, Meng-Jie 513
 Lin, Jianhua 47
 Lin, Li 513
 Lin, Xinqian 789
 Liou, Cheng-Yuan 14
 Liou, Daw-Ran 14
 Liu, Binxiang 868
 Liu, Feng 418
 Liu, GuoJun 385
 Liu, Guoqi 136
 Liu, Handiang 505
 Liu, Heping 47
 Liu, Hui 748
 Liu, Jiafeng 427
 Liu, Jianzhuang 844
 Liu, Jin 176, 230
 Liu, Jiwei 880
 Liu, Ju 797
 Liu, Juan 844
 Liu, Lei 880
 Liu, Ming 649
 Liu, Tianliang 168
 Liu, Xiabi 54
 Liu, Yi 279, 481
 Liu, Ying 705
 Liu, Zhi 827
 Long, Zhiliang 554
 Lu, Wen 184
 Lu, Wenjun 208
 Lu, Xinguo 410
 Lu, Ying 6
 Luo, Jiebo 168
 Ma, Bo 30
 Ma, Hao 731
 Ma, Wenguang 78
 Ma, Xinlu 6
 Ma, Yu 497
 Maitimusha, Kuerban 40
 Mei, Ning 184
 Meng, Qingjie 359
 Mi, Jian-Xun 820
 Mo, Yiming 168
 Ni, Weiping 6
 Pandit, Sagar 22
 Pei, Mingtao 30
 Peng, Yuanfan 143
 Qian, Chengshan 642, 691
 Qin, Shiyin 376, 649
 Qu, Bingxin 94
 Qu, Yan-Yun 513
 Ren, Dongwei 287
 Ren, Mingwu 611
 Shen, Haojie 246, 852
 Shen, Yuanyuan 312
 Sheng, Jingwei 554
 Shi, Chen 143
 Shi, Dong yan 401
 Shi, Meng-Yuan 714
 Shi, Xin 47

- Shi, Yongchang 208
 Shi, Zhihong 279
 Simak, Alex A. 14
 Song, Guojie 546
 Song, Ruizhuo 618
 Song, Xu 595
 Sun, Bing 465
 Sun, Changyin 1, 344, 642, 691, 731,
 764, 875, 880
 Sun, Jiande 797
 Tang, Jingsheng 295, 368
 Tang, Kezong 868
 Tang, XiangLong 385, 427, 538
 Tang, Zhen-Min 666
 Tao, Dacheng 336, 351, 697
 Teng, Peng 54
 Tong, Chao-nan 579
 Tong, Xiaomin 78
 Tu, Yi-Cheng 22
 Wang, Bin 697
 Wang, Bing 103
 Wang, Chang-Dong 86, 112
 Wang, Chengyun 489
 Wang, Guoping 465
 Wang, Hanzi 312, 320, 441
 Wang, Hong-Qiang 820
 Wang, Hong-Yuan 561
 Wang, Huan 611
 Wang, Jing Ni 626
 Wang, Le 303
 Wang, Li 271
 Wang, Liang 418, 603
 Wang, Lichao 835
 Wang, Lina 376
 Wang, Lisheng 497
 Wang, Nan 336
 Wang, Qianping 176, 230
 Wang, Ruiping 200
 Wang, Wei 731, 764
 Wang, Xiaoming 263
 Wang, Xiumei 570
 Wang, Yilun 554, 813
 Wang, Ying 570
 Wang, Yuxiang 47
 Wang, Zhengpeng 70
 Wei, Qinglai 618
 Wei, Wei 359
 Wen, Bo 579
 Weng, Shifeng 682
 Wu, Guorong 554
 Wu, Jianhui 1
 Wu, Jianzhai 216
 Wu, Junzheng 6
 Wu, Ou 449
 Wu, Xihong 473, 481
 Wu, Yuwei 30
 Xia, Yu 603
 Xiang, Jinlong 255
 Xiao, Wendong 618
 Xiao, Xiuchun 86
 Xie, Jiangchuan 797
 Xie, Kunqing 546
 Xing, Junliang 152
 Xu, Jiajie 457
 Xu, Jindong 160
 Xu, Jun 401
 Xu, Lai 529
 Xu, Lei 887
 Xu, Xin 393, 722
 Xu, Yong 457
 Xu, Zhi Wei 626
 Xue, Lei 642
 Xue, Song 208
 Yan, Dong 312, 320
 Yan, Hui 635
 Yan, Weidong 6
 Yan, Yan 312, 320, 441
 Yang, Jian 223, 457
 Yang, Jinfeng 120, 136, 152, 449
 Yang, Min 30
 Yang, Nana 40
 Yang, Shunqing 192
 Yang, Tao 78, 94, 771
 Yang, Wankou 344, 418, 875
 Yang, Xi 748
 Yang, Xu 579
 Yang, Yi 835
 Yang, Yuning 473
 Ye, Lei 722
 Ye, Xinfu 344, 764
 Yin, Erwei 295, 368
 Yin, Yixin 70, 756
 Yu, Sufen 587
 Yu, Xianchuan 160
 Yu, Xiaotian 62
 Yu, Yang 295, 368

- Yuan, Lin 128
Yuan, Yan 805
Zeng, Yan 579
Zhan, De-Chuan 714
Zhang, Baochang 844
Zhang, Changshui 682
Zhang, Congcong 827
Zhang, David 287, 328
Zhang, Fang-Chao 103
Zhang, Guoyun 1
Zhang, Haigang 756
Zhang, Hai-Ying 513
Zhang, Han 6
Zhang, Hong 827
Zhang, Hongzhi 287, 328, 789
Zhang, Ji 561
Zhang, Jiang 554
Zhang, Jingmei 642, 691
Zhang, Lan 103
Zhang, Lefei 351
Zhang, Lei 359
Zhang, Liangpei 336, 351
Zhang, Ruimin 642, 691
Zhang, Sen 756
Zhang, Weiping 880
Zhang, Yanning 78, 94, 359, 497, 587,
595, 860
Zhang, Yingtao 538
Zhang, Yunong 62
Zhao, Feng 505
Zhao, Jia 868
Zhao, Qijun 827
Zhao, Yingdi 143
Zhao, Zhenbing 303
Zheng, Junjie 813
Zheng, Rui 393
Zhou, Li 128
Zhou, Shan 152
Zhou, Yan-Wen 513
Zhou, Yue 658, 672, 779
Zhou, Zongtan 295, 368
Zhu, Lian 344, 764
Zhu, Limin 658
Zhu, Qi 457
Zhu, Xingquan 22
Zhu, Xiuchang 168
Zhuo, Li 143, 246, 852
Zou, Chuhang 238
Zuo, Lei 393, 722
Zuo, Wangmeng 287, 328, 789