

De-Shuang Huang
Vitoantonio Bevilacqua
Juan Carlos Figueroa
Prashan Premaratne (Eds.)

LNCS 7995

Intelligent Computing Theories

9th International Conference, ICIC 2013
Nanning, China, July 2013
Proceedings

 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

De-Shuang Huang
Vitoantonio Bevilacqua
Juan Carlos Figueroa
Prashan Premaratne (Eds.)

Intelligent Computing Theories

9th International Conference, ICIC 2013
Nanning, China, July 28-31, 2013
Proceedings

Volume Editors

De-Shuang Huang
Tongji University, Machine Learning and Systems Biology Laboratory
Shanghai 201804, China
E-mail: dshuang@tongji.edu.cn

Vitoantonio Bevilacqua
Polytechnic of Bari, Electrical and Electronics Department
70125 Bari, Italy
E-mail: vitoantonio.bevilacqua@gmail.com

Juan Carlos Figueroa
District University Francisco José de Caldas, Faculty of Engineering
Bogotá, Colombia
E-mail: jcfigueroag@udistrital.edu.co

Prashan Premaratne
University of Wollongong, School of Electrical, Computer and
Telecommunications Engineering, North Wollongong, NSW 2522, Australia
E-mail: prashan@uow.edu.au

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-39478-2 e-ISBN 978-3-642-39479-9
DOI 10.1007/978-3-642-39479-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013942311

CR Subject Classification (1998): I.2, I.4, I.5, H.4, J.3, F.1, F.2, C.2, H.3, H.2.8,
G.2.2, K.6.5

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

© 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.

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

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

ICIC 2013, held in Nanning, China, July 28–31, 2013, constituted the 9th International Conference on Intelligent Computing. It built upon the success of ICIC 2012, ICIC 2011, ICIC 2010, ICIC 2009, ICIC 2008, ICIC 2007, ICIC 2006, and ICIC 2005 that were held in Huangshan, Zhengzhou, Changsha, China, Ulsan, Korea, Shanghai, Qingdao, Kunming, and Hefei, China, respectively.

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

ICIC 2013 received 561 submissions from 27 countries and regions. All papers went through a rigorous peer-review procedure and each paper received at least three review reports. Based on the review reports, the Program Committee finally selected 192 high-quality papers for presentation at ICIC 2013, included in three volumes of proceedings published by Springer: one volume of *Lecture Notes in Computer Science* (LNCS), one volume of *Lecture Notes in Artificial Intelligence* (LNAI), and one volume of *Communications in Computer and Information Science* (CCIS).

This volume of *Lecture Notes in Computer Science* (LNCS) includes 74 papers.

The organizers of ICIC 2013, including Tongji University and Guangxi University for Nationalities, made an enormous effort to ensure the success of the conference. We hereby would like to thank the members of the Program Committee and the referees for their collective effort in reviewing and soliciting the papers. We would like to thank Alfred Hofmann, executive editor at Springer, for his frank and helpful advice and guidance throughout and for his continuous support in publishing the proceedings. In particular, we would like to thank all the authors for contributing their papers. Without the high-quality submissions from the authors, the success of the conference would not have been possible.

Finally, we are especially grateful to the IEEE Computational Intelligence Society, the International Neural Network Society, and the National Science Foundation of China for their sponsorship.

May 2013

De-Shuang Huang
Vitoantonio Bevilacqua
Juan Carlos Figueroa
Prashan Premaratne

ICIC 2013 Organization

General Co-chairs

De-Shuang Huang, China
Marios Polycarpou, Cyprus
Jin-Zhao Wu, China

Program Committee Co-chairs

Kang-Hyun Jo, Korea
Pei-Chann Chang, Taiwan, China

Organizing Committee Co-chairs

Yong-Quan Zhou, China
Bing Wang, China

Award Committee Co-chairs

Laurent Heutte, France
Phalguni Gupta, India

Publication Chair

Juan Carlos Figueroa, Colombia

Workshop/Special Session Chair

Vitoantonio Bevilacqua, Italy

Special Issue Chair

Michael Gromiha, India

Tutorial Chair

Luonan Chen, Japan

International Liaison Chair

Prashan Premaratne, Australia

Publicity Co-chairs

Kyungsook Han, Korea
Lei Zhang, China
Ling Wang, China
Valeriya Gribova, Russia

Exhibition Chair

Xing-Ming Zhao, China

Organizing Committee Members

Yong Huang, China
Yong Wang, China
Yuanbin Mo, China

Conference Secretary

Su-Ping Deng, China

Program Committee Members

Andrea Francesco Abate, Italy	Xiyuan Chen, China
Vasily Aristarkhov, Russian Federation	Yang Chen, China
Costin Badica, Romania	Michal Choras, Poland
Soumya Banerjee, India	Angelo Ciaramella, Italy
Waqas Haider Khan Bangyal, Pakistan	Jose Alfredo F. Costa, Brazil
Vitoantonio Bevilacqua, Italy	Mingcong Deng, Japan
Shuhui Bi, China	Eng. Salvatore Distefano, Italy
Zhiming Cai, Macau	Mariagrazia Dotoli, Italy
Chin-Chih Chang, Taiwan, China	Haibin Duan, China
Pei-Chann Chang, Taiwan, China	Hazem Elbakry, Egypt
Guanling Chen, USA	Karim Faez, Iran
Luonan Chen, Japan	Jianbo Fan, China
Jingdong Chen, China	Jianwen Fang, USA
Songcan Chen, China	Minrui Fei, China
Weidong Chen, China	Juan Carlos Figueroa, Colombia

Wai-Keung Fung, Canada
 Jun-Ying Gan, China
 Liang Gao, China
 Xiao-Zhi Gao, Finland
 Dunwei Gong, China
 Valeriya Gribova, Russia
 M. Michael Gromiha, India
 Xingsheng Gu, China
 Kayhan Gulez, Turkey
 Phalguni Gupta, India
 Fei Han, China
 Kyungsook Han, Korea
 Yong-Tao Hao, China
 Jim Harkin, UK
 Haibo He, USA
 Jing Selena He, USA
 Laurent Heutte, France
 Wei-Chiang Hong, Taiwan, China
 Yuexian Hou, China
 Heyan Huang, China
 Kun Huang, USA
 Zhenkun Huang, China
 Bora Peter Hung, Ireland
 Chuleerat Jaruskulchai, Thailand
 Umarani Jayaraman, India
 Li Jia, China
 Zhenran Jiang, China
 Kang-Hyun Jo, Korea
 Dong-Joong Kang, Korea
 Sanggil Kang, Korea
 Muhammad Khurram Khan,
 Saudi Arabia
 Donald H. Kraft, USA
 Harshit Kumar, Korea
 Yoshinori Kuno, Japan
 Takashi Kuremoto, Japan
 Vincent C S Lee, Australia
 Bo Li, China
 Guo-Zheng Li, China
 Kang Li, UK
 Min Li, China
 Shi-Hua Li, China
 Xiaou Li, Mexico
 Honghuang Lin, USA
 Chunmei Liu, USA
 Ju Liu, China
 Ke Lv, China
 Jinwen Ma, China
 Lorenzo Magnani, Italy
 Xiandong Meng, USA
 Tarik Veli Mumcu, Turkey
 Roman Neruda, Czech Republic
 Ken Nguyen, USA
 Ben Niu, China
 Yusuke Nojima, Japan
 Sim-Heng Ong, Singapore
 Francesco Pappalardo, Italy
 Young B. Park, Korea
 Surya Prakash, India
 Prashan Premaratne, Australia
 Seeja K.R., India
 Ajita Rattani, Italy
 Ivan Vladimir Meza Ruiz, Mexico
 Angel D. Sappa, Spain
 Li Shang, China
 Fanhuai Shi, China
 Jiatao Song, China
 Stefano Squartini, Italy
 Zhan-Li Sun, China
 Evi Syukur, Australia
 Naoyuki Tsuruta, Japan
 Antonio E. Uva, Italy
 Katya Rodriguez Vazquez, Mexico
 Jun Wan, USA
 Bing Wang, China
 Lei Wang, China
 Ling Wang, China
 Shitong Wang, China
 Wei Wang, China
 Yijie Wang, China
 Wei Wei, China
 Zhi Wei, China
 Qiong Wu, China
 Xiaojun Wu, China
 Yan Wu, China
 Junfeng Xia, China
 Shunren Xia, China
 Yuanqing Xia, China
 Liangjun Xie, USA
 Bingji Xu, China

Hua Xu, USA
 Shao Xu, Singapore
 Zhenyu Xuan, USA
 Tao Ye, China
 Wen Yu, Mexico
 Boyun Zhang, China
 Lei Zhang, HongKong, China
 Xiang Zhang, USA
 Yi Zhang, China

Hongyong Zhao, China
 Xing-Ming Zhao, China
 Zhongming Zhao, USA
 Bo-Jin Zheng, China
 Chun-Hou Zheng, China
 Fengfeng Zhou, China
 Shuigeng Zhou, China
 Li Zhuo, China

Reviewers

Marjan Abdechiri	Jair Cervantes	Sheng Ding
Aliahmed Adam	Hyunuk Chae	Shihong Ding
Erum Afzal	Aravindan Chandrabose	Xiang Ding
Sabooah Ajaz	James Chang	Joaquín Dopazo
Felix Albu	Yuchou Chang	Vlad Dovgalecs
Muhammad Amjad	Chun Chen	Vladislavs Dovgalecs
Deepa Anand	David Chen	Guangyue Du
Mary Thangakani	Diyi Chen	Ji-Xiang Du
Anthony	Gang Chen	Haibin Duan
Vasily Aristarkhov	Jianhung Chen	Qiqi Duan
Sepehr Attarchi	Songcan Chen	Saber Elsayed
Amelia Badica	Chi-Tai Cheng	Kadir Erkan
Leemon Baird	Cong Cheng	Villatoro-Tello Esaú
Abdullah Bal	Ferdinando Chiacchio	Mahdi Ezoji
Waqas Bangyal	Cheng-Hsiung Chiang	Shaojing Fan
Donato Barone	Shen Chong	Yaping Fang
Ye Bei	Angelo Ciaramella	Chen Fei
Olivier Berder	Azis Ciayadi	Chong Feng
Simon Bernard	Rudy Ciayadi	Liangbing Feng
Vitoantonio Bevilacqua	Mike Collins	Alessio Ferone
Shuhui Bi	Danilo Comminiello	Francesco Ferrise
Jun Bo	Carlos Cubaque	Juan Carlos Figueroa
Nora Boumella	Yan Cui	Michele Fiorentino
Marius Brezovan	Dajundu	Qian Fu
Fabio Bruno	Francesca De Crescenzo	Hironobu Fujiyoshi
Fanliang Bu	Kaushik Deb	Wai-keung Fung
Ni Bu	Sara Dellantonio	Liang Gao
Guorong Cai	Jing Deng	Xiaofang Gao
Qiao Cai	Lei Deng	Yang Gao
Francesco Camastra	Suping Deng	Yushu Gao
Giuseppe Carbone	Somnath Dey	Zhong-Ke Gao
Raffaele Carli	Liya Ding	Dingfei Ge

Jing Ge	Shouling Ji	Juan Li
Giorgio Gemignani	Zhiwei Ji	keling Li
Shaho Ghanei	Hongjun Jia	Qingfeng Li
Saameh Golzadeh	Changan Jiang	Qinghua Li
Jing Gu	He Jiang	Shang Li
Smile Gu	Min Jiang	Wei Li
Tower Gu	Shujuan Jiang	Xiangyang Li
Xingsheng Gu	Ying Jiang	Xiaodi Li
Jian Guan	Yizhang Jiang	Xiaoguang Li
Shi-Jie Guan	Yunsheng Jiang	Yunqi Li
Lanshen Guo	Lie Jie	Jing Liang
Tiantai Guo	Xu Jie	Xinwu Liang
Weili Guo	Ning-De Jin	Gumei Lin
Yinan Guo	Wei Jin	Jian Lin
Puneet Gupta	Mingyuan Jiu	Yong Lin
Haciilhan	Ren Jun	Chenbin Liu
Javad Haddadnia	Yang Kai	Chih-Chin Liu
Fei Han	Hee-Jun Kang	Huai-Jen Liu
Kyungsook Han	Olesya Kazakova	James Liu
Meng Han	Ondrej Kazik	Jin-Xing Liu
Wenting Han	Mohebbi Keyvan	Li Liu
Yu-Yan Han	Amar Khoukhi	Liangxu Liu
Xin Hao	Hong-hyun Kim	Qing Liu
Manabu Hashimoto	One-Cue Kim	Xiaoming Liu
Selena He	Taeho Kim	Yijian Liu
Tao He	Wooyoung Kim	Yufeng Liu
German Hernandez	Ogaard Kirk	Yuhang Liu
Laurent Heutte	Duangmalai Klongdee	Zhe Liu
Anush Himanshu	Kunikazu Kobayashi	Alfredo Liverani
Huabin Hong	Toshiaki Kondo	Francesco Longo
Lei Hou	Kitti Koonsanit	SiowYong Low
Changjun Hu	Takashi Kuremoto	Xingjia Lu
Ke Hu	Baeguen Kwon	Zhen Lu
Haoqian Huang	Hebert Lacey	Junfeng Luo
Huali Huang	Qixun Lan	Durak-Ata Lutfiye
Jida Huang	Jose A. Fernandez Leon	Jun Lv
Ke Huang	Bingnan Li	Chuang Ma
Lei Huang	Bo Li	Lan Ma
Yea-Shung Huang	Chen Li	Wencai Ma
Wu-Yin Hui	Dalong Li	Xiaotu Ma
Sorin Ilie	Fuhai Li	Xiaoxiao Ma
Saiful Islam	Hui Li	Shingo Mabu
Saeed Jafarzadeh	Jianqing Li	Sakashi Maeda
Chuleerat Jaruskulchai	Jianxing Li	Mohammad-Javad
James Jayaputera	Jingfei Li	Mahmoodabadi

Guoqin Mai	Stefanos Quartini	Jie Sun
Swanirbhar Majumder	Muhammad Rahman	Jing Sun
Mario Manzo	Sakthivel Ramasamy	Sheng Sun
Antonio Maratea	Muhammad Ramzan	Xiaoyan Sun
Erik Marchi	Tao Ran	Yu Sun
Hunny Mehrotra	Muhammad Rashid	Jayasudha John Suseela
Geethan Mendiz	Hamidreza Rashidy	Lijing Tan
Giovanni Merlino	Kanan	Buzhou Tang
Hyeon-Gyu Min	Abdul Rauf	Xinhua Tang
Saleh Mirheidari	Angelo Riccio	Xiwei Tang
Akio Miyazaki	Lisbeth Rodríguez	Tansalg
Raffaele Montella	Sudha Sadasivam	Zhu Teng
Tsuyoshi Morimoto	Angelo Antonio Salatino	Hongjun Tian
Saeed Mozaffari	Angel Sappa	Tian Tian
Lijun Mu	Michele Scarpiniti	DungLe Tien
Tarik Veli Mumcu	Donguk Seo	Aruna Tiwari
Francesca Nardone	Chao Shao	Kamlesh Tiwari
Ken Nguyen	Haojie Shen	Mukesh Tiwari
Zhen Ni	Yehu Shen	Minglei Tong
Changhai Nie	Bo Sheng	Ximo Torres
Aditya Nigam	Fanhuai Shi	Joaquín Torres-Sospedra
Zhijun Niu	Jibin Shi	Farzad Towhidkhah
Ryuzo Okada	Xiutao Shi	Yao-Hong Tsai
Kazunori Onoguchi	Atsushi Shimada	Naoyuki Tsuruta
Dazhao Pan	Nobutaka Shimada	Gurkan Tuna
Quanke Pan	Ye Shuang	Pierpaolo Valentini
Jekang Park	Jakub Smid	Andrey Vavilin
Anoosha Paruchuri	Jakub Smidbjunior	Tomaso Vecchi
Giangluca Percoco	Jakub Smidmff	Giuseppe Vettigli
Alfredo Pereira	Mai Son	Petra Vidnerová
Elisano Pessa	Bin Song	Aihui Wang
Fausto Petrella	Rui Song	Bin Wang
Martin Pilat	Yang Song	Chun-Hsin Wang
Gibran-Fuentes Pineda	Yinglei Song	Fang-Fang Wang
Surya Prakash	Jairo Soriano	Haili Wang
Prashan Premaratne	Sotanto Sotanto	Huisen Wang
Miguel A. Pujana	Stefano Squartini	Jingchuan Wang
Kang Qi	Antonino Staiano	Jinhe Wang
Xiangbo Qi	Hung-Chi Su	Jun Wang
Pengjiang Qian	Jinya Su	Ling Wang
Kaijin Qiu	Rina Su	Mingyi Wang
Ying Qiu	Eng.Marco Suma	Qixin Wang
Chenghua Qu	Marco Suma	Sheng-Yao Wang
Junfeng Qu	Guangming Sun	Shulin Wang
Junjun Qu	Jiankun Sun	Xiangyu Wang

Xiao Wang	Yu Xue	Qiangfeng Zhang
Xiaoming Wang	Atsushi Yamashita	Ruofei Zhang
Xiying Wang	Mingyuan Yan	Shuyi Zhang
Yan Wang	Yan Yan	Wenxi Zhang
Yichen Wang	Chia-Luen Yang	Xianxia Zhang
Yong Wang	Chyuan-Huei Yang	Xiaoling Zhang
Yongcui Wang	Shan-Xiu Yang	Xiujun Zhang
Yunfei Wang	Wankou Yang	Yanfeng Zhang
Zhaoxi Wang	Wenqiang Yang	Yifeng Zhang
Zi Wang	Yang Yang	Yong-Wei Zhang
Zongyue Wang	Altshuler Yaniv	Zhanpeng Zhang
Suparta Wayan	Xiangjuan Yao	Changbo Zhao
Wei Wei	Shin Yatakahashi	Guodong Zhao
Zhijia Wei	Tao Ye	Liang Zhao
Zhixuan Wei	Myeong-Jae Yi	Miaomiao Zhao
Ouyang Wen	Kai Yin	Min Zhao
Shengjun Wen	Hua Yu	Xinhua Zhao
Chao Wu	Liu Yu	Xu Zhao
Hongrun Wu	Wu Yu	Yue Zhao
Jingli Wu	Jinghua Yuan	Yunlong Zhao
Weili Wu	Lin Yuan	Bojin Zheng
Yonghui Wu	Quan Yuan	Chunhou Zheng
Qing Xia	Assunta Zanetti	Huanyu Zheng
Siyu Xia	Samir Zeghlache	Min Zheng
Qin Xiao	Yu Zeng	Xiaolong Zheng
Yongfei Xiao	Zhiyong Zeng	Xinna Zheng
Keming Xie	Chunhui Zhang	Liugui Zhong
Minzhu Xie	Chunjiang Zhang	Jiayin Zhou
Zhenping Xie	Duwen Zhang	Linhua Zhou
Chao Xing	Guanglan Zhang	Songsheng Zhou
Wei Xiong	Guohui Zhang	Yinzhi Zhou
Dawen Xu	Hailei Zhang	Hua Zhu
Jin Xu	Hongyun Zhang	Nanli Zhu
Jing Xu	Jianhua Zhang	Xuefen Zhu
Xiaoyin Xu	Jing Zhang	Yongxu Zhu
Xin Xu	Jun Zhang	Zhongjie Zhu
Ye Xu	Kevin Zhang	Majid Ziaratban
Yuan Xu	Ming Zhang	
Zhenyu Xuan	Peng Zhang	

Table of Contents

Neural Networks

Research on the Application of T-S Fuzzy Neural Network in Quantitative Identification of Mixed Gas	1
<i>Yu Zhang</i>	

Nature Inspired Computing and Optimization

A Self-adaptive Hybrid Population-based Incremental Learning Algorithm for M -Machine Reentrant Permutation Flow-shop Scheduling	8
<i>Zuo-Cheng Li, Bin Qian, Rong Hu, Chang-Sheng Zhang, and Kun Li</i>	

A Model of Emotional Intelligent Agent for Cooperative Goal Exploration	21
<i>Takashi Kuremoto, Tetsuya Tsurusaki, Kunikazu Kobayashi, Shingo Mabu, and Masanao Obayashi</i>	

Cognitive Science and Computational Neuroscience

The Cloud Model Based on Grey System Theory and Application on Effectiveness Evaluation	31
<i>Yanbin Shi, Hong Liu, and Jian Sun</i>	

Knowledge Discovery and Data Mining

Finding the Optimal Bus-Routes Based on Data Mining Method	39
<i>Yong Wang, Yang Liu, Cheng-zhi Zhang, and Zhi-ping Li</i>	

Analysis and Application of Data Mining in CRM System of Some Mobile Branch	47
<i>Song-Tao Lou, Ji-Rui Li, and Wen-Bin Zheng</i>	

Line Clipping Algorithm of Affine Transformation for Polygon	55
<i>Wenjun Huang</i>	

Evaluating Community Detection Using a Bi-objective Optimization . . .	61
<i>Nesrine Ben Yahia, Narjès Bellamine Ben Saoud, and Henda Ben Ghezala</i>	

Evolutionary Learning and Genetic Algorithms

New Algorithm for Solving Nonlinear Equations Roots	71
<i>Delong Guo, Yongquan Zhou, and Xiaobin Luo</i>	
Relationship between Blackwell's Informativeness and Supermodular Precision	77
<i>Yuanyuan Li</i>	
A Novel Genetic Programming Based Classifier Design Using a New Constructive Crossover Operator with a Local Search Technique	86
<i>Arpit Bhardwaj and Aruna Tiwari</i>	

Machine Learning Theory and Methods

Forecasting Method of Stock Price Based on Polynomial Smooth Twin Support Vector Regression	96
<i>Shifei Ding, HuaJuan Huang, and Ru Nie</i>	
ApLeafis: An Android-Based Plant Leaf Identification System	106
<i>Lin-Hai Ma, Zhong-Qiu Zhao, and Jing Wang</i>	

Natural Language Processing and Computational Linguistics

Microblog Searching Module Based on Community Detection	112
<i>Pengxiang Lin, Tingting He, and Yong Zhang</i>	

Fuzzy Theory and Models

Ambiguity Preferences and Games	120
<i>Lan Sun</i>	
Fuzzy Decision Making Based on Fuzzy Propositional Logic with Three Kinds of Negation	128
<i>Zhenghua Pan</i>	
The Strong Perron Integral of Fuzzy Number Valued Functions	141
<i>Yabin Shao, Huanhuan Zhang, Yongchun Cao, and Yuqin Bai</i>	
Similarity of Fuzzy Triangular Number Based on Indifference Area and Its Application	150
<i>Xixiang Zhang and Jianxun Liu</i>	

Fuzzy Systems and Soft Computing

Research on Image Mosaic Algorithm Based on Computer Wizard Vector Field Algorithm	156
<i>Xiaobo Gao and Xianmei Fang</i>	
Conditional Value at Risk Methodology under Fuzzy-Stochastic Approach	163
<i>Shao-fang Tang and Ying-yu He</i>	
Application of the Fuzzy Set with Three Kinds of Negation FSCOM in the Stock Investment	173
<i>Jiexin Zhao and Zhenghua Pan</i>	

Unsupervised and Reinforcement Learning

Two-Phase Image Segmentation with the Competitive Learning Based Chan-Vese (CLCV) Model	183
<i>Yanqiao Zhu, Anhui Wang, and Jinwen Ma</i>	
Adaptive Step Searching for Solving Stochastic Point Location Problem	192
<i>Tongtong Tao, Hao Ge, Guixian Cai, and Shenghong Li</i>	

Intelligent Computing in Finance/Banking

Application of GSO Algorithm to the Parameter Estimation of Option Pricing Model	199
<i>Yuanbin Mo, Fuyong Liu, and Yanzhui Ma</i>	

Intelligent Computing in Communication Networks

Heuristics for Hub Location Problems with Alternative Capacity Levels and Allocation Constraints	207
<i>Jeng-Fung Chen</i>	
A Multi-objective Genetic Algorithm Based Handoff Decision Scheme with ABC Supported	217
<i>Chengbo Zhang, Xingwei Wang, and Min Huang</i>	

Intelligent Computing in Petri Nets/Transportation Systems

A Weight-Based Graph Coloring Approach to Airport Gate Assignment Problem	227
<i>Yanjun Jiang and Xueyan Song</i>	

Intelligent Computing in Network Software/Hardware

Game-Based Scheduling Algorithm to Achieve Optimize Profit in MapReduce Environment	234
<i>Cong Wan, Cuirong Wang, Ying Yuan, and Haiming Wang</i>	
Research and Application of Regression Test Cases Aided Design Method Based on Association-Mode	241
<i>Wenhong Liu, Xin Wu, Yuheng Hao, and Jimao Wang</i>	
Virtual Network Embedding Algorithm Based Connective Degree and Comprehensive Capacity	250
<i>Ying Yuan, Cuirong Wang, Na Zhu, Cong Wan, and Cong Wang</i>	

Intelligent Control and Automation

Fire Prevention in Gas Stove with Gas Sensors	259
<i>Soojung Kang, Bitna Song, and Ilhoon Shin</i>	

Intelligent Image/Document Retrievals

Word Spotting Application in Historical Mongolian Document Images	265
<i>Hongxi Wei and Guanglai Gao</i>	

Intelligent Data Fusion and Information Security

Binary Program Statistical Features Hiding through Huffman Obfuscated Coding	275
<i>Xiaopeng Niu, Qingbao Li, Wei Wang, and Xiaokang Weng</i>	

Intelligent Prediction and Time Series Analysis

Glaucus: Predicting Computing-Intensive Program's Performance for Cloud Customers	285
<i>Xia Liu, Zhigang Zhou, Xiaojiang Du, Hongli Zhang, and Junchao Wu</i>	
A Multiwavelet Support Vector Machine Prediction Algorithm for Avionics PHM	295
<i>Xin Zhou, Zheng Xiang, Meng Liu, and Jiang Xiang</i>	

Intelligent Sensor Networks

- Using a Real-Time Top-k Algorithm to Mine the Most Frequent Items
over Multiple Streams 305
Ling Wang, Zhao Yang Qu, Tie Hua Zhou, and Keun Ho Ryu

Intelligent Fault Diagnosis

- Sequential Diagnosis Method for Rotating Machinery Using Support
Vector Machines and Possibility Theory 315
Hongtao Xue, Ke Li, Huaqing Wang, and Peng Chen

Knowledge Representation/Reasoning and Expert Systems

- On M-Type Bag Structures 325
Kankana Chakrabarty

Virtual Reality and Human-Computer Interaction

- Coupled Tissue Bleeding Simulation in Virtual Surgery 331
*Cheng Yang, Jiaxiang Guo, Jie Han, Xiangyun Liao, and
Zhiyong Yuan*

Intelligent Computing in Signal Processing

- The Study of Adaptive Modulation Algorithms in OFDM System 339
Yanli Shi, Yanbin Shi, Renxia Ou, and Haixia Yu
- An Improved Algorithm of Parameters Estimation for Frequency-
Hopping Signal 346
Jun Lv, Weitao Sun, and Tong Li
- The Feature Extraction Method of EEG Signals Based on Degree
Distribution of Complex Networks from Nonlinear Time Series 354
Fenglin Wang, Qingfang Meng, Weidong Zhou, and Shanshan Chen
- Threshold Estimation Method for Spectrum Sensing Using Bootstrap
Technique 362
Liping Luo, Wei Zhou, and Huazhi Meng

Intelligent Computing in Pattern Recognition

- Palm Print Feature Extraction and Recognition Based on BEMD-ICA
II and LS-SVM 368
Gui-Ping Dai

Detection and Classification of Defect Patterns in Optical Inspection Using Support Vector Machines	376
<i>Liangjun Xie, Rui Huang, and Zhiqiang Cao</i>	
Illumination Invariant Face Recognition	385
<i>Guangyi Chen, Sridhar Krishnan, Yongjia Zhao, and Wenfang Xie</i>	
Evaluation of Sampling Methods for Learning from Imbalanced Data . . .	392
<i>Garima Goel, Liam Maguire, Yuhua Li, and Sean McLoone</i>	
A Genealogical Study of the Origin of Pashtuns	402
<i>Hameed Ullah Khan and Nasir Ahmed</i>	
Two Notes from Experimental Study on Image Steganalysis	411
<i>Qingxiao Guan, Jing Dong, and Tieniu Tan</i>	

Intelligent Computing in Biometrics Recognition

Robust and Efficient Iris Recognition Based on Sparse Error Correction Model	421
<i>Wei Cao, Yun Song, Zunliang He, and Zhimin Zhou</i>	
Face Verification across Age Progressing Based on Active Appearance Model and Gradient Orientation Pyramid	427
<i>Xing Wu, Ji-Xiang Du, and Chuan-Min Zhai</i>	
Single Sample Face Recognition Based on DCT and Local Gabor Binary Pattern Histogram	435
<i>Zhihua Xie</i>	
Iris Classification Based on Its Quality	443
<i>Aditya Nigam, Anvesh T., and Phalguni Gupta</i>	
Age-Invariant Face Recognition Using Shape Transformation	453
<i>Shubham Jain, Aditya Nigam, and Phalguni Gupta</i>	

Intelligent Computing in Image Processing

Low-Resolution Image Restoration Using the Combination Method of Sparse Representation and PDE Model	462
<i>Li Shang and Zhan-li Sun</i>	
An Improved Image Corner Matching Approach	472
<i>Bijin Yan, Fanhuai Shi, and Jiguang Yue</i>	
A Visual Dataflow Model for the Process Flow of Remote Sensing Products	482
<i>Bing Zhou, Guan-feng Wu, Yong Xu, Jia-guo Li, and Yang Liu</i>	

Image Denoising Algorithm Based on Edge-Preserving Self-Snake Model and Wavelet-Based PDE	490
<i>Changxiong Zhou, Shufen Lui, Tingqin Yan, and Wenlin Tao</i>	
Single Sample Face Recognition Based on Multiple Features and Twice Classification	498
<i>Xiaohua Wang, Wei Liu, Min Hu, and Liangfeng Xu</i>	
A Method of Alteration Information Extracted of Uranium Mine Base on TM Remote Sensing Image	507
<i>Lin Jiang, Yuefei Yi, Shengyu You, Zhibo Wang, and Linlin He</i>	
A Novel Multi-Scale Local Region Model for Segmenting Image with Intensity Inhomogeneity	516
<i>Xiao-Feng Wang, Hai Min, and Yi-Gang Zhang</i>	

Intelligent Computing in Robotics

Fault Tolerant Control for Robot Manipulators Using Neural Network and Second-Order Sliding Mode Observer	526
<i>Mien Van and Hee-Jun Kang</i>	
Fusion of Vision and Inertial Sensors for Position-Based Visual Servoing of a Robot Manipulator	536
<i>Tran Minh Duc and Hee-Jun Kang</i>	

Intelligent Computing in Computer Vision

Efficient 3D Reconstruction for Urban Scenes	546
<i>Weichao Fu, Lin Zhang, Hongyu Li, Xinfeng Zhang, and Di Wu</i>	
Combining Edge and One-Point RANSAC Algorithm to Estimate Visual Odometry	556
<i>Van-Dung Hoang, Danilo Cáceres Hernández, and Kang-Hyun Jo</i>	
Automatic Reconstruction of Two-Dimensional Broken Objects	566
<i>Yanjuan Zhu</i>	

Special Session on Biometrics System and Security for Intelligent Computing

An Efficient Natural Image Deblurring Algorithm	576
<i>Rajesh R. Pillai, Vandana Dixit Kaushik, and Phalguni Gupta</i>	
A Heuristic Technique for Performance Improvement of Fingerprint Based Integrated Biometric System	584
<i>Kamlesh Tiwari, Soumya Mandi, and Phalguni Gupta</i>	

Edge Based Steganography on Colored Images 593
Mangat Rai Modi, Saiful Islam, and Phalguni Gupta

Special Session on Intelligent Computing and Personalized Assisted Living

An Evolutionary Optimization Method for Parameter Search in 3D Points Cloud Reconstruction 601
Vitoantonio Bevilacqua, Fabio Ivona, Domenico Cafarchia, and Francescomaria Marino

Computer Human Interaction Using Multiple Visual Cues and Intelligent Computing

Hexahedral Mesh Generation for Geometry with Multi-featured Constraints 612
Xing Dai, Han-Guo Cui, Li-Ping Zhang, Zheng-Min Li, and Fei-Zhang Wang

Complex Systems Theory and Methods

Computing Model of Individual Emotion in the Mass Incidents with Venting Anger 621
Fanliang Bu and Yiyi Wang

A SVM-Based System for Predicting Protein-Protein Interactions Using a Novel Representation of Protein Sequences 629
Zhuhong You, Zhong Ming, Ben Niu, Suping Deng, and Zeruan Zhu

Exploiting SAaaS in Smart City Scenarios 638
Salvatore Distefano, Giovanni Merlino, and Antonio Puliafito

Author Index 649

Research on the Application of T-S Fuzzy Neural Network in Quantitative Identification of Mixed Gas

Yu Zhang

School of Electronic Information Engineering, Suzhou Vocational University,
Suzhou International Educational Park, Road 106, Zhineng Dadao,
Suzhou 215104, Jiangsu, China
zhixin331@163.com

Abstract. This article integrated rule expression capacity of fuzzy logic inference with self-learning ability of the neural network, proposed to build Takagi-Sugeno fuzzy neural network's quantitative identification of mixed gas by combining T-S fuzzy neural network with neural network. The results indicated that this system has generalization, learning, mapping capabilities. It can better realize quantitative identification of mixed gas. This system will provide method for intelligent identification of mixed gas.

Keywords: T-S fuzzy neural network, gas analysis, quantitative identification.

1 Introduction

In recent years, car ownership in China is growing. Meanwhile, car accidents are also rising. Car accidents caused by drunk driving triggered most casualties and financial loss. Therefore, in order to prevent drunk driving, it's valuable to detect the concentration of alcohol to determine whether the drivers are drunk driving. One of the key factors is the dynamic characteristics information processing technology of gas sensor array [1] [2]. Compared with traditional quantitative-model identification, this system is more accurate in measuring alcohol concentration in mixed gas on the basis of T-S fuzzy neural network[3-5].

2 T-S Fuzzy Neural Network System Structures

The network consists of two sub-networks: the condition neural network and the conclusion neural network. The condition neural network provides each fuzzy rule with correspondent adaptability; the conclusion neural network calculates the outcome of fuzzy rules. The final outcome of network is the weighted sum of all rules. Weighting coefficient is the adaptability of each fuzzy rule output by condition neural network. From the perspective of overall function, this is a fuzzy system. From the perspective of network topological structure, it is a neural network consists of

nodes and weights. Therefore, fuzzy neural network has both advantages of fuzzy system and neural network. It has strong mapping and adaptive ability.

(1) Conditions portion neural network

The condition neural network consists of four layers. The first layer is the input layer; each node is directly connected with the respective components x_i of the input vector. Then transfer the input value $X = [x_1, x_2, \dots, x_n]^T$ to the next layer. The nodes of the layer $N_1 = n$

Each node of the second layer represents a language variable. Its role is to calculate subordinating degree function μ_i^j which belongs to fuzzy sets of different languages, among which:

$$\mu_i^j = \mu_{A_i^j}(x_i) \quad i = 1, 2, \dots, n, j = 1, 2, \dots, m_i \quad (1)$$

n is input dimension, m_i is the fuzzy segmentation number of x_i . In the research in this article, subordinating degree function is using gaussian function

$$\mu_i^j = \exp\left(-\frac{x_i - c_{ij}}{\sigma_{ij}}\right)^2 \quad (2)$$

Among which, c_{ij} and σ_{ij} represent the center and width of subordinating degree

function. This layer's node amount is $N_2 = \sum_{i=1}^n m_i$.

Each node of the third layer represents a fuzzy rule; its role is to calculate adaptability of every rule. That is:

$$\alpha_j = \min\{\mu_1^{i1}, \mu_2^{i2}, \dots, \mu_n^{in}\} \quad (3)$$

Among which,

$$i_1 \in \{1, 2, \dots, m_1\} \quad i_n \in \{1, 2, \dots, m_n\}, j = 1, 2, \dots, m, m = \prod_{i=1}^n m_i \quad (4)$$

The node amount of this layer is $N_3 = m$. As for the given input, only language variable near input point can have large subordinating degree function. When subordinating degree function is small, it approaches 0. Therefore, when there are

fewer nodes input instead of 0 in α_j while most node input is 0, this is similar to local approximation network.

The nodes in fourth layer are same with the third layer, which is $N_4 = N_3 = m$, it realizes normalization computing, that is:

$$\overline{\alpha_j} = \frac{\alpha_j}{\sum_{j=1}^m \alpha_j}, j = 1, 2, \dots, m \tag{5}$$

(2) conclusion neural network

The subsequent network consists of r parallel sub-networks with the same structure; each sub-network can have an input. The first layer of sub-network is input layer. It will transfer the variable to the second layer. The 0 node of input layer $x_0 = 1$, its role is to provide constant term output by fuzzy rules. That is:

$$y_{ij} = p_{j0}^i + p_{j1}^i x_1 + \dots + p_{jn}^i x_n = \sum_{k=0}^n p_{jk}^i x_k, i = 1, 2, \dots, r, j = 1, 2, \dots, m \tag{6}$$

The third layer of sub-network is the output by computer system, that is:

$$y_i = \sum_{j=1}^m \overline{\alpha_j} y_{ij} (i = 1, 2, \dots, r) \tag{7}$$

Thus, y_i is the weight sum of each rule output, weighting coefficient is the normalization adaptability of fuzzy rules, which is the third layer link weight of condition neural network's output as conclusion neural network.

3 Systematic Learning Algorithms

Assuming the fuzzy segmentation number of each component input is predefined, learning parameters are mainly connection weight of conclusion neural network $p_{jk}^i (j = 1, 2, \dots, m; k = 0, 1, \dots, n; i = 1, 2, \dots, r)$, and the central value of subordinate function in the second layer of condition neural network $\sigma_{ij} (i = 1, 2, \dots, m; j = 1, 2, \dots, m_i)$.

This article chooses error cost function:

$$E = \frac{1}{2} \sum_{i=1}^r (t_i - y_i)^2 \tag{8}$$

In this formula, t_i and y_i represent desired output and actual output. First, the learning algorithm of parameter p_{jk}^i is:

$$\frac{\partial E}{\partial p_{jk}^i} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial p_{jk}^i} = -(t_i - y_i) \overline{\alpha_j} x_k \quad (9)$$

$$p_{jk}^i(l+1) = p_{jk}^i(l) - \beta \frac{\partial E}{\partial p_{jk}^i} = p_{jk}^i(l) + \beta(t_i - y_i) \overline{\alpha_j} x_k \quad (10)$$

In this formula, $j = 1, 2, \dots, m; k = 0, 1, \dots, n; i = 1, 2, \dots, r$

On this basis, when talking about learning problem of c_{ij} and σ_{ij} , the p_{jk}^i can be fixed. At this time, consequent y_{ij} after every rule changed to connection weight w_{ij} in the last layer. In the output layer, this article chooses the weighted average method to realize the specified calculation, that is:

$$y_i = \sum_{j=1}^m \overline{\alpha_j} y_{ij} \quad (i = 1, 2, \dots, r) \quad (11)$$

Error cost function is $E = \frac{1}{2} \sum_{i=1}^r (t_i - y_i)^2$, by means of the error back propagation algorithm to calculate

$$\delta_i^{(5)} = (t_i - y_i), i = 1, 2, \dots, n \quad (12)$$

$$\delta_i^{(4)} = \sum_{j=1}^r \delta_i^{(5)} y_{ij}, j = 1, 2, \dots, m \quad (13)$$

$$\delta_i^{(3)} = \delta_i^{(4)} \frac{\sum_{i=1, j \neq i}^m \alpha_i}{(\sum_{i=1}^m \alpha_i)^2}, j = 1, 2, \dots, m \quad (14)$$

$$\delta_i^{(2)} = \sum_{j=1}^r \delta_i^{(3)} s_{ij} e^{\frac{(x_i - c_{ij})}{\sigma_{ij}^2}}, i = 1, 2, \dots, n, j = 1, 2, \dots, m \quad (15)$$

When *and* adopted minimum operation and μ_i^j is the minimum of the k th rule node, $s_{ij} = 1$, otherwise, $s_{ij} = 0$. Obtain:

$$\frac{\partial E}{\partial c_{ij}} = -\delta_{ij}^{(2)} \frac{2(x_i - c_{ij})}{\delta_{ij}^2} \tag{16}$$

$$\frac{\partial E}{\partial \delta_{ij}} = -\delta_{ij}^{(2)} \frac{2(x_i - c_{ij})^2}{\delta_{ij}^3} \tag{17}$$

$$c_{ij}(k+1) = c_{ij}(k) - \beta \frac{\partial E}{\partial c_{ij}} \tag{18}$$

$$\delta_{ij}(k+1) = \delta_{ij}(k) - \beta \frac{\partial E}{\partial \delta_{ij}} \tag{19}$$

In this formula, $\beta > 0$, it is learning rate: $i = 1, 2, \dots, n, j = 1, 2, \dots, m_i$

4 Results and Analysis of System Identification

8 sensors were used in the experiment, including one temperature, one humidity, and one smoke sensor. Apart from that, there are 5 heater-type SnO_2 gas sensitive sensors. The alcohol concentration was in the range of 0 ~ 300ppm, the gasoline ranges from 0 to 270 ppm. The temperature ranges from 20 to 35 ° C, humidity ranges from 20 ~ 80% RH with proper smoke concentration. Air adopted conventional static injection method. Detect the response features of sensor array without gas interference and with gasoline interference. 8 sensor response values is set as a sample under each mixed gas condition. Choose 48 groups of different interference state and randomly divided them into training and testing samples.

There are 3 input neurons and 5 output neurons in T-S fuzzy neural network , indicating the alcohol concentration. Determine 3 linguistic variable values for each input layer neurons: $m_1 = m_2 = m_3 = 3$,

Use function *genfis1* to generate a fuzzy inference system by mesh segmentation based on a given data. The input and output subordinating degree function curve of this fuzzy inference system were evenly divided while make sure it covers whole input and output interval. Subordinating degree function used gaussian function. The System Quantitative results of the identification section and the relative error is shown in Table 1.

Table 1. Quantitative result of the identification section and the relative error

Num	Gas interfere	gas(ppm)	alcohol(ppm)	output(ppm)	relative error
1	without	90	100	100.0726	0.0726
2	without	90	200	199.9371	0.0315
3	without	90	300	299.6252	0.1249
4	without	180	100	100.0314	0.0314
5	without	180	200	200.0201	0.1010
6	without	180	300	299.9801	0.0064
7	without	270	100	99.8111	0.1889
8	without	270	200	199.7291	0.1355
9	without	270	300	299.8994	0.0335
10	have	90	100	99.9840	0.0160
11	have	90	200	199.4688	0.2656
12	have	90	300	299.2120	0.2630
13	have	180	100	100.1495	0.1495
14	have	180	200	200.1880	0.0940
15	have	180	300	300.2598	0.0866
16	have	270	100	99.7459	0.2541
17	have	270	200	200.6111	0.3056
18	have	270	300	300.2670	0.0890

From Table 1, we can see that the biggest error of system quantitative identification is 0.3875%, the overall average error is 0.1059%, the actual error is less than 10 ppm. As indicated in table 1, Quantitative identification result is in accordance with actual concentration, showing that T-S fuzzy neural network can better have the quantitative identify alcohol concentration.

5 Conclusions

Through the above analysis we can see: combining the expression ability of fuzzy logic inference rules and neural network self-learning ability, using T-S fuzzy neural network to cope with signal adopted by air sensor array, we can better realize the quantitative identification of mixed gas. This will provide new method for intelligent identification of mixed gas.

References

1. Wang, S.T.: Fuzzy systems, fuzzy neural network and application design. Science and Technology Literature Publishing House, Shanghai (1998)
2. Wen, X., Zhou, L.: Matlab neural network simulation and application. Science Press, Beijing (2011)
3. Wang, Y.N.: Intelligent information processing technology. Higher Education Press, Beijing (2009)
4. Qu, J.L., Wang, L., Gao, F.: The use of artificial neural networks for quantitative analysis of the mixed gas. Northwestern Polytechnical University 19(3), 399–402 (2010)
5. Zhou, J.H., Wang, L., Yao, C.H.: T-S fuzzy neural network based on fuzzy neural network. Journal of Tsinghua University (Science and Technology) (37), 76–80 (2008)

A Self-adaptive Hybrid Population-Based Incremental Learning Algorithm for M -Machine Reentrant Permutation Flow-Shop Scheduling

Zuo-Cheng Li, Bin Qian^{*}, Rong Hu, Chang-Sheng Zhang, and Kun Li

Department of Automation, Kunming University of Science and Technology,
Kunming 650500, China
bin.qian@vip.163.com

Abstract. This paper proposes a self-adaptive hybrid population-based incremental learning algorithm (SHPBIL) for the m -machine reentrant permutation flow-shop scheduling problem (MRPFSSP) with makespan criterion. At the initial phase of SHPBIL, the information entropy (IE) of the initial population and an *Interchange*-based search are utilized to guarantee a good distribution of the initial population in the solution space, and a training strategy is designed to help the probability matrix to accumulate information from the initial population. In SHPBIL's global exploration, the IE of the probability matrix at each generation is used to evaluate the evolutionary degree, and then the learning rate is adaptively adjusted according to the current value of IE, which is helpful in guiding the search to more promising regions. Moreover, a mutation mechanism for the probability model is developed to drive the search to quite different regions. In addition, to enhance the local exploitation ability of SHPBIL, a local search based on critical path is presented to execute the search in some narrow and promising search regions. Simulation experiments and comparisons demonstrate the effectiveness of the proposed SHPBIL.

Keywords: Population-based incremental learning algorithm, m -machine reentrant permutation flow-shop scheduling, information entropy, self-adaptive strategy, critical-path-based local search.

1 Introduction

This paper deals with the m -machine reentrant permutation flow-shop scheduling problem (MRPFSSP) with makespan criterion, which represents a typical subsystem in some semiconductor manufacturing system [1]. In MRPFSSP with n jobs and m machines, each job should be processed on machine $1, 2, \dots, m, 1, 2, \dots, m, \dots, 1, 2, \dots, m$, and the job sequence is the same for each machine. It has been proved that the MRPFSSP is NP-hard, even for the two-machine case [2]. Choi et al. [1] presented several types of heuristic algorithms for the MRPFSSP to minimize makespan.

^{*} Corresponding author.

Chen et al. [3] proposed a hybrid genetic algorithm (HGA) for the reentrant flow-shop scheduling problem (RFSSP) with makespan criterion. Chen et al. [4] developed a HGA to obtain approximate optimal solutions for the reentrant permutation flow-shop scheduling problem (RPFSSP) with makespan criterion. RPFSSP can be classified as a special case of MRPFSSP, in which the sequence of jobs is the same for each machine at each level. Therefore, some approaches for RPFSSP also can be expanded to address MRPFSSP.

As a novel probabilistic-model based evolutionary algorithm, population-based incremental learning algorithm (PBIL) was introduced by Baluja [5] for solving traveling salesman problem (TSP) and job-shop scheduling problem. PBIL uses a variable-independence probability model to generate new population and guide the search direction, which is the earliest model of estimation of distribution algorithm (EDA). The evolution process of PBIL is regarded as a process of competitive learning, and its probability model is updated by the current best solution at each generation to accumulate the information of excellent individuals. Due to its simple framework and outstanding global exploration ability, PBIL has attracted much attention and has been used to solve some production scheduling problems. Salhi et al. [6] proposed a PBIL to deal with hybrid flow-shop scheduling problem. Wang et al. [7] developed a hybrid PBIL for flexible job-shop scheduling problem, whose local search scheme is designed based on the critical path. Pang et al. [8] presented an adaptive PBIL algorithm for solving job-shop and flow-shop scheduling problem, which can obtain satisfied solutions.

In this paper, we propose a self-adaptive hybrid PBIL (SHPBIL) for MRPFSSP. Firstly, the information entropy (IE) of the initial population and an *Interchange*-based search are utilized to guarantee a good distribution of the initial population in the solution space, and a training strategy is designed to help the probability matrix to accumulate information from the initial population. Secondly, in SHPBIL's global exploration, the IE of the probability matrix at each generation is used to evaluate the evolutionary level, and then the learning rate is adaptively adjusted according to the current value of IE, which is helpful in guiding the search to more promising regions. Thirdly, a mutation mechanism for the probability model is developed to drive the search to quite different regions. Fourthly, to enhance the local exploitation ability of SHPBIL, a local search based on critical path is presented to execute the search in some narrow and promising search regions. Due to the hybridization of global search and local search, MRPFSSP can be solved effectively. Simulation experiments and comparisons show the effectiveness and robustness of the proposed SHPBIL.

The remainder of this paper is organized as follows. In the next section, the MRPFSSP is stated briefly. In section 3, SHPBIL is presented and described in detail. In section 4, simulation result and comparisons are given. Finally, we end the paper with some conclusion and future work in section 5.

2 Problem Description

The problem size is denoted by $n \times m \times L$ where n is the number of jobs, m is the number of machines, and L means the repeat or reentrant times. The MRPFSSP is

usually defined as follows. Each job should be processed on machine $1, 2, \dots, m, 1, 2, \dots, m, \dots, 1, 2, \dots, m$. In this case, every job can be decomposed into several layers and each layer starts on 1 and finishes on. That is, each job should visits every machine L ($L > 1$) times. The job ordering is the same on all machines. No preemption is allowed. Once an operation is started, it must be completed without any interruption. At any time, each machine can process at most one job.

Denote $\pi = [\pi_1, \pi_2, \dots, \pi_{n \times L}]$ the sequence or schedule of jobs to be processed, π_j ($j \in \{1, \dots, n \times L\}$) the j th job in π , $l(\pi_j)$ the repeat or reentrant times of job π_j in $[\pi_1, \dots, \pi_j]$, $p(\pi_j, k, l(\pi_j))$ the $l(\pi_j)$ th processing time of job π_j on machine k , $C(\pi_j, k, l(\pi_j))$ the $l(\pi_j)$ th completion time of job π_j on machine k . Let $C(\pi_0, k, l(\pi_0)) = 0$ and $C(\pi_j, k, 0) = 0$ for $k \in \{1, \dots, m\}$ and $j \in \{1, \dots, n \times L\}$. Then $C(\pi_j, k, l(\pi_j))$ can be calculated as follows:

$$C(\pi_j, 1, l(\pi_j)) = \max\{C(\pi_{j-1}, 1, l(\pi_{j-1})), C(\pi_j, m, l(\pi_j) - 1)\} + p(\pi_j, 1, l(\pi_j)), j = 1, \dots, n \times L, \quad (1)$$

$$C(\pi_j, k, l(\pi_j)) = \max\{C(\pi_{j-1}, k, l(\pi_{j-1})), C(\pi_j, k - 1, l(\pi_j))\} + p(\pi_j, k, l(\pi_j)), j = 1, \dots, n \times L, k = 2, \dots, m. \quad (2)$$

Thus, the makespan can be defined as

$$C_{\max}(\pi) = C(\pi_{n \times L}, m, l(\pi_{n \times L})) \quad (3)$$

The MRPFSSP with the makespan criterion is to find a schedule π^* in the set of all schedules Π such that

$$\pi^* = \arg\{C_{\max}(\pi)\} \rightarrow \min, \quad \forall \pi \in \Pi. \quad (4)$$

Obviously, the size of Π is $(n \times L)!/[L!]^n$.

3 SHPBIL

3.1 Solution Representation

Based on the properties of MRPFSSP, we adopt the operation-based solution representation, that is, every individual of the population is a feasible solution of the MRPFSSP, for example, $[\pi_1, \pi_2, \dots, \pi_{4 \times 2}] = [1, 2, 4, 2, 1, 3, 3, 4]$ is an individual when the problem's scale $n \times m \times L$ is set to $4 \times 3 \times 2$.

3.2 Probability Model and Updating Mechanism

Different from other evolutionary algorithms, PBIL generates new population by sampling a probability model (i.e., probability matrix). Hence, the probability model

has a key effect on the performances of the PBIL. In this study, the probability matrix is defined as follows:

$$P_{matrix}(gen) = \begin{pmatrix} P_{11}(gen) & \dots & P_{1C}(gen) \\ \vdots & \ddots & \vdots \\ P_{n1}(gen) & \dots & P_{nC}(gen) \end{pmatrix}_{n \times C}, \quad (5)$$

where $C = n \times L$, $\sum_{w=1}^n P_{wj}(gen) = 1$, and $P_{wj}(gen)$ is the probability of job w appearing in the j th position of π at generation gen .

Let $\pi_i(gen) = [\pi_{i1}(gen), \pi_{i2}(gen), \dots, \pi_{iC}(gen)]$ denote the i th individual in SHPBIL's population at generation gen , and $\alpha(gen)$ the learning rate at generation gen . The matrix $P_{matrix}(gen+1)$ is updated according to $\pi_i(gen)$ by the following two steps:

- Step 1: Set $x = \pi_{ij}(gen)$ and $p_{xj}(gen+1) = p_{xj}(gen) + \alpha(gen)$ for $j = 1, \dots, n \times L$.
- Step 2: Set $p_{wj}(gen+1) = p_{wj}(gen+1) / \sum_{y=1}^n p_{yj}(gen+1)$ for $w = 1, \dots, n$ and $j = 1, \dots, n \times L$.

3.3 New Population Generation Method

In each generation of the SHPBIL, the new individuals are generated by sampling the probability matrix mentioned in 3.2. Denote PS the size of the population and $SelectJob(\pi_i(gen+1), j)$ the function of selecting a job in the j th position of $\pi_i(gen+1)$ by using the matrix $P_{matrix}(gen)$. The procedure of $SelectJob(\pi_i(gen+1), j)$ is described as follows:

- Step 1: Randomly create a probability r where $r \sim [0, 1)$.
- Step 2: Get a candidate job CJ by the roulette-wheel selection scheme.
- Step 2.1: If $r \sim [0, p_{1j}(gen))$, then set $CJ = 1$ and go to Step 3.
- Step 2.2: If $r \sim [\sum_{y=1}^w p_{yj}(gen), \sum_{y=1}^{w+1} p_{yj}(gen))$ and $w \in \{1, \dots, n-1\}$, then set $CJ = w+1$ and go to Step 3.
- Step 3: Return CJ .

Let $l(CJ, \pi_i(gen+1))$ denote the repeat times of CJ in $\pi_i(gen+1)$. Then, the new population generation method is given as the following steps:

- Step 1: Set $i = 1$.
- Step 2: Generate a new individual $\pi_i(gen+1)$.
- Step 2.1: Set $\pi_{ij}(gen+1) = 0$ for $j = 1, \dots, n$.
- Step 2.2: Set $j = 1$.

Step 2.3: $CJ = \text{SelectJob}(\pi_i(\text{gen} + 1), j)$.

Step 2.4: If $l(CJ, \pi_i(\text{gen} + 1)) = L$, then go to Step 2.3.

Step 2.5: Set $\pi_{ij}(\text{gen} + 1) = CJ$.

Step 2.6: Set $j = j + 1$.

Step 2.7: If $j \leq n \times L$, then go to Step 2.3.

Step 3: Set $i = i + 1$.

Step 4: If $i \leq PS$, then go to Step 2.

3.4 Population Initialization

The SHPBIL produces the initial population by using information entropy [9] and *Interchange*-based neighborhood search. Firstly, we generate the population based on information entropy theory to guarantee the initial population has a better distribution in the solution space. Secondly, since *Interchange* is a simple and effective neighborhood search in the existing literatures, an *Interchange*-based search can be utilized to enhance the quality of individuals in the population. According to our previous tests, we found that the initial population generated by our method could avoid both “super” and “similar” individuals effectively and the performance of our method was better than random-based or heuristic-based method.

Denote $pop(\text{gen})$ the population at generation gen , $IE_j(\text{gen})$ the information entropy of the j th position of all individuals in $pop(\text{gen})$, $IE_pop(\text{gen})$ the information entropy of $pop(\text{gen})$, and IE_0_pop the threshold of $IE_pop(\text{gen})$. $IE_j(\text{gen})$ can be calculated by directly using the method in [9]. $IE_pop(\text{gen})$ can be expressed as:

$$IE_pop(\text{gen}) = \sum_{j=1}^{n \times L} IE_j(\text{gen}) / (n \times L). \quad (6)$$

We generate the initial population $pop(0)$ by the following steps:

Step 1: Set IE_0_pop to a certain value.

Step 2: Set $pop(0) = \text{null}$.

Step 3: Generate the first individual $\pi_1(0)$ randomly and add it to $pop(0)$.

Step 4: $i = 2$.

Step 4.1: Generate an individual $\pi_i(0)$ randomly and add it to $pop(0)$.

Step 4.2: Calculate $IE_pop(0)$.

Step 4.3: If $IE_pop(0) \leq IE_0_pop$, then delete $\pi_i(0)$ from $pop(0)$.

Step 4.4: $i = i + 1$.

Step 4.5: If $i \leq PS$, then go to Step 4.1.

Step 5: Output $pop(0)$.

Denote $Interchange(\pi_i(gen), u, v)$ the interchange of the job at the u th position (i.e., $\pi_{iu}(gen)$) and the job at the v th position (i.e., $\pi_{iv}(gen)$). Then, the final $pop(0)$ can be obtained by using the *Interchange*-based neighborhood search as follows:

Step 0: Set $i=1$.

Step 1: Randomly choose two positions u and v , where $\pi_{iu}(0) \neq \pi_{iv}(0)$. Then, set

$$\pi_i^1(0) = Interchange(\pi_i(0), u, v).$$

Step 2: Set $LOOP = 1$.

Step 2.1: Randomly choose two positions u and v , where $\pi_{iu}^1(0) \neq \pi_{iv}^1(0)$.

$$\text{Then, set } \pi_i^2(0) = Interchange(\pi_i^1(0), u, v).$$

Step 2.2: If $f(\pi_i^2(0)) < f(\pi_i^1(0))$, then set $\pi_i^1(0) = \pi_i^2(0)$.

Step 2.3: $LOOP = LOOP + 1$.

Step 2.4: If $LOOP < n \times L \times (n \times L - 1) / 2$, then go to Step 2.1.

Step 3: If $f(\pi_i^1(0)) < f(\pi_i(0))$, then set $\pi_i(0) = \pi_i^1(0)$.

Step 4: Set $i=i+1$. If $i \leq PS$, then goto Step 1.

Step 5: Output the final $pop(0)$.

3.5 Probability Matrix Training

The population initialization strategy in the above subsection can guarantee the initial population has a better distribution in the solution space. However, based on our previous tests, the PBIL algorithm with the proposed population initialization strategy was still likely to fall into low-quality local optima at the start phase. This phenomenon shows the fact that the probability matrix $P_{matrix}(gen)$ does not have enough historical information to track the relatively high-quality individuals in the initial population. Therefore, an information-entropy-based probability matrix training method is designed to enhance PBIL's search ability. Denote IE_{tra_ini} the initial training information entropy, $IE_{tra_P_{matrix}}(0)$ the information entropy of $P_{matrix}(0)$, K_{tra} ($K_{tra} < 1$) the training constant, IE_{tra_0} the threshold of $IE_{tra_P_{matrix}}(0)$, and $\pi^{lbest}(0) = [\pi_1^{lbest}(0), \dots, \pi_{n \times L}^{lbest}(0)]$ the best individual of the initial population $pop(0)$. Based on the method in [8], $IE_{tra_P_{matrix}}(0)$ can be calculated by the following formulation:

$$IE_{tra_P_{matrix}}(0) = -\sum_{i=1}^n \sum_{j=1}^{n \times L} P_{ij}(0) \times \ln(P_{ij}(0)). \quad (7)$$

The procedure of probability matrix training can be expressed as follows:

Step 0: Set $IE_{tra_ini} = n \times L \times \ln(n)$.

Step 1: Set $IE_{tra_0} = K_{tra} \times IE_{tra_ini}$.

Step 2: Train the probability matrix.

Step 2.1: Randomly select u and v , where $\pi_u^{lbest}(0) \neq \pi_v^{lbest}(0)$.

Step 2.2: $\pi_{neighbor}^{lbest}(0) = Interchange(\pi^{lbest}(0), u, v)$.

Step 2.3: Update the probability matrix $P_{matrix}(0)$ according to $\pi_{neighbor}^{lbest}(0)$ by using the updating method in subsection 3.2.

Step 2.4: Calculate $IE_{tra} - P_{matrix}(0)$ of the current $P_{matrix}(0)$ by using (7).

Step 2.5: If $IE_{tra} - P_{matrix}(0) > IE_{tra_0}$, then go to Step 2.1.

Step3: Output a new probability matrix $P_{matrix}(0)$ and the current $IE_{tra} - P_{matrix}(0)$.

It can be seen from the above procedure that Step 2.1 and Step 2.2 compose a perturbation operator, and Step 2.5 is used to help $P_{matrix}(0)$ to accumulate information from the *Interchange*-based neighbors of the best individual in $pop(0)$. Moreover, the output $IE_{tra} - P_{matrix}(0)$ is used in the calculation of a threshold IE_0 in the next subsection.

3.6 Self-adaptive Adjusting Strategy

PBIL is a stochastic search algorithm, which guides the search direction by sampling the probability model. The updating mechanism of probability model is a key factor to perform global exploration. However, in the process of probability model updating, learning rate is a sensitive parameter. That is, a small value of learning rate will slow down the convergence speed. On the contrary, a large value of learning rate will cause the population to converge too early. Thus, it is important to choose suitable value for learning rate.

The information entropy is utilized to measure the evolutionary degree. The evolution of PBIL is a process of accumulating the excellent solutions' historical information. That is, during the evolutionary process of PBIL, each $P_{wj}(gen)$ in probability matrix $P_{matrix}(gen)$ changes gradually from initial values to 0 or 1. Let $IE - P_{matrix}(gen)$ denote the information entropy of $P_{matrix}(gen)$. By using the method in [8], $IE - P_{matrix}(gen)$ can be expressed as:

$$IE - P_{matrix}(gen) = -\sum_{i=1}^n \sum_{j=1}^{n \times L} P_{ij}(gen) \times \ln(P_{ij}(gen)), \quad gen \geq 0. \quad (8)$$

Obviously, $IE - P_{matrix}(gen)$ decreases with an increase of gen , and $IE - P_{matrix}(gen)$ tends to 0 when gen increases to a large value. Denote $IE_0 = k_0 \times IE_{tra} - P_{matrix}(0)$ the threshold of $IE - P_{matrix}(gen)$ and α_0 the maximum learning rate. The learning rate $\alpha(gen)$ is adaptively adjusted according to the formula as follows:

$$\alpha(gen) = \begin{cases} \exp[k_1 \times (IE - P_{matrix}(gen)/IE_0 - 1)] \times \alpha_0 & IE - P_{matrix}(gen) < IE_0 \\ \beta^{k_2 \times (IE - P_{matrix}(gen)/IE_0 - 1)} \times \alpha_0 & IE - P_{matrix}(gen) \geq IE_0 \end{cases}, \quad (9)$$

where k_0 , k_1 , and k_2 are the adjusting parameters and β ($0 < \beta < 1$) is the base number of exponential function. The self-adaptive adjusting strategy is reflected in formula (9). That is, with the increase of gen , $\alpha(gen)$ is set to a relatively small value to increase the population diversity and track the population at the beginning phase, and it is set to a comparatively large value to speed up the process of accumulating information at the middle phase, and it is set to a small value to enhance the search precision and avoid falling into local optima at the last phase.

3.7 Mutation Mechanism

Each $P_{wj}(gen)$ in probability matrix $P_{matrix}(gen)$ changes gradually from initial values to 0 or 1 when gen increases to a large value. This means the population diversity and exploration ability are decreased with the increase of gen . Inspired by Bajula's work [10], we propose a mutation operator to disturb the probability matrix. Denote $\alpha_m(gen)$ the mutation rate at generation gen and set $\alpha_m(gen) = \alpha(gen) / 2$. The process of the proposed mutation mechanism can be described as follows:

Step 1: Randomly select u , v and q , where $u \neq v \neq q$.

Step 2: Randomly select ω_1 , ω_2 and ω_3 .

Step 3: Set $P_{\omega_1 u}(gen) = P_{\omega_1 u}(gen) + \alpha_m(gen)$, $P_{\omega_2 v}(gen) = P_{\omega_2 v}(gen) + \alpha_m(gen)$, and $P_{\omega_3 q}(gen) = P_{\omega_3 q}(gen) + \alpha_m(gen)$.

Step 4: For $w = 1, \dots, n$, set $P_{wu}(gen) = P_{wu}(gen) / \sum_{y=1}^n P_{yu}(gen)$, $P_{wv}(gen) = P_{wv}(gen) / \sum_{y=1}^n P_{yv}(gen)$, and $P_{wq}(gen) = P_{wq}(gen) / \sum_{y=1}^n P_{yq}(gen)$.

3.8 Critical-Path-Based Local Search

3.8.1 Critical Path and Block

Referring to the work of Grabowski and Wodecki [11], the MRPFSSP can also be described by a graph model, which illustrates the technological constraints for each job and the processing sequence on each machine. The longest path from node $(1, 1)$ to $(m, n \times L)$ in the graph is defined as the critical path, and the length of the critical path is equal to C_{\max} .

3.8.2 Search Strategy

Grabowski and Wodecki [11] had given a detailed definition of the moves and neighborhood structure for the permutation flow-shop problem, which can avoid invalid moves inside blocks. We extend Grabowski's structure in this subsection, which can also provide some narrow and promising search regions. Denote $Insert(\pi_i, u, v)$ the insertion of the job π_{iu} before π_{iv} when $u > v$ and after π_{iv} when $u < v$,

$B_k = [\pi_k^1, \pi_k^2, \dots, \pi_k^{kb}]$ the k th block and kb the length of B_k . Then, we define *Left moves* $M_L(B_k)$ and *Right moves* $M_R(B_k)$ as follows:

$M_L(B_k)$: *Insert*($B_k, u, 1$) for $u = 2, \dots, kb$.

$M_R(B_k)$: *Insert*(B_k, u, kb) for $u = 1, \dots, kb-1$.

Denote $\pi^{lbest}(gen) = [\pi_1^{lbest}(gen), \pi_2^{lbest}(gen), \dots, \pi_{n \times L}^{lbest}(gen)]$ the local best individual of the current population $pop(gen)$, BL the total number of blocks in $\pi^{lbest}(gen)$, $N^{lbest}(B_k)$ the set of neighbors of $\pi^{lbest}(gen)$ when applying $M_L(B_k)$ and $M_R(B_k)$ to $\pi^{lbest}(gen)$, and $FindBestN^{lbest}(\pi^{lbest}(gen))$ the scanning procedure of finding the best neighbor in $N^{lbest}(B_1) \cup N^{lbest}(B_2) \cup \dots \cup N^{lbest}(B_{BL})$. The procedure of the critical-path-based local search for $\pi^{lbest}(gen)$ is given as follows:

Step 1: Set $\pi^{i-0} = \pi^{lbest}(gen)$.

Step 2: *Insert*-based perturbation.

Step 2.1: Randomly select u and v , where $\pi_u^{lbest}(gen) \neq \pi_v^{lbest}(gen)$ and $u > v$.

Step 2.2: $\pi^{lbest}(gen) = \text{Insert}(\pi^{lbest}(gen), u, v)$.

Step 3: $\pi^{i-1} = \text{FindBestN}^{lbest}(\pi^{lbest}(gen))$.

Step 4: If $f(\pi^{i-1}) < f(\pi^{i-0})$, then $\pi^{i-0} = \pi^{i-1}$.

Step 5: Output π^{i-0} .

3.9 Procedure of SHPBIL

Based on the contents in the above subsections, we propose the procedure of SHPBIL as follows:

Step 0: Denote π^{gbest} the global best individual and $genMax$ the maximum generation.

Step 1: Initialization.

Step 1.1: Set $gen = 0$.

Step 1.2: Generate the initial population $pop(0)$ by using the method in subsection 3.4 and set $\pi^{gbest} = \pi^{lbest}(0)$.

Step 1.3: Set $P_{wj}(gen) = 1/n$ for $w = 1, \dots, n$ and $j = 1, \dots, n \times L$.

Step 1.4: Train the probability matrix $P_{matrix}(0)$ by using the method in subsection 3.5.

Step 2: Set $gen = gen + 1$.

Step 3: Calculate the learning rate $\alpha(gen-1)$ by using the self-adaptive adjusting strategy in subsection 3.6.

- Step 4: Generate the population $pop(gen)$ by using the new population generation method in subsection 3.3, and calculate the makespan of each individual and update $\pi^{lbest}(gen)$.
- Step 5: Disturb the probability matrix $P_{matrix}(gen-1)$ by using the mutation mechanism in subsection 3.7.
- Step 6: Apply critical-path-based local search in subsection 3.8 to $\pi^{lbest}(gen)$ and update π^{gbest} .
- Step 7: Update the probability matrix $P_{matrix}(gen)$ according to π^{gbest} by using the updating mechanism in subsection 3.2.
- Step 8: If $gen < genMax$, then go to Step2.
- Step 9: Output π^{gbest} .

It can be seen from the above procedure that Step 7 uses π^{gbest} obtained in Step 6 and the self-adaptive $\alpha(gen-1)$ obtained in Step 3 to update the probability matrix, which means new generated individuals can aptly absorb the information of the global best individual during the evolution process and then guide the search to more promising regions, and Step 5 is the perturbation operator, which can restrain the search from dropping into local optima and drive the search to quite different regions. Moreover, Step 6 performs exploitation from the regions obtained by Step 4. Since both exploration and exploitation are stressed and balanced, SHPBIL is hopeful to obtain good results.

4 Simulation Result and Comparisons

4.1 Experimental Design

In order to test the performance of the proposed SHPBIL, a set of instances under different scales is randomly generated. The $n \times m \times L$ combinations include $10 \times 5 \times 3$, $10 \times 8 \times 6$, $20 \times 10 \times 3$, $20 \times 12 \times 5$, $30 \times 10 \times 3$, $30 \times 15 \times 6$, $40 \times 10 \times 2$, $40 \times 15 \times 3$, $50 \times 10 \times 3$, and $50 \times 20 \times 4$. The processing time $p(\pi_j, k, l(\pi_j))$ is generated from a uniform distribution [1, 100]. All algorithms are coded in Delphi7.0 and are executed on Mobile Intel Core 2 Duo 2.0 GHz processor with 2GB memory.

For each instance, each algorithm is run 20 times independently. Based on our previous experiments, the parameters of SHPBIL are set as follows: the population size $PS = 50$, the threshold $IE_0_pop = 0.2$, the maximum learning rate $\alpha_0 = 0.04$, the adjusting parameters $k_0 = 0.9$, $k_1 = 0.3$ and $k_2 = 0.7$, the base number of exponential function $\beta = 0.25$ and the training constant $K_{tra} = 0.25$.

4.2 Comparisons of HGA, Random+MN3+SO, HGA_V, and SHPBIL

For the purpose of showing the effectiveness of SHPBIL, we compare SHPBIL with a hybrid genetic algorithm (HGA) [4] and a so-called Random+MN3+SO algorithm

[1]. HGA is an effective algorithm for RPFSSP. Random+MN3+SO algorithm is one of the most effective algorithms for MRPFSPP [1]. Moreover, we also compare SHPBIL with HGA_V, which is a variant of an effective HGA for RFSSP [3]. In HGA_V, we generate the first part of initial population randomly. The maximum generations of HGA, HGA_V and SHPBIL are set to 1000, 50000 and 500, respectively. The running time of Random+MN3+SO is decided only by the scale of problem. The simulation results are listed in Table 1 and Table 2, where *BEST* denotes the best makespan, *AVG* denotes the average makespan, *WORST* denotes the worst makespan, T_{avg} denotes the average running time, and *SD* denotes the standard derivation.

From Table 1 and Table 2, it is shown that the SHPBIL is better than HGA, HGA_V and Random+MN3+SO with respect to solution quality. The values of *AVG*,

Table 1. Comparisons of *BEST* and *AVG* of HGA, HGA_V, Random+MN3+SO and SHPBIL

Instances	HGA		HGA_V		Random+MN3+SO		SHPBIL		
	<i>BEST</i>	<i>AVG</i>	<i>BEST</i>	<i>AVG</i>	<i>BEST</i>	<i>AVG</i>	<i>BEST</i>	<i>AVG</i>	<i>WORST</i>
10×5×3	1995	2014.10	1990	2029.45	2186	2323.40	1968	1995.70	2012
10×8×6	4265	4325.25	4595	4672.00	4986	5367.35	4247	4287.75	4399
20×10×3	3914	3960.60	3877	3933.35	3976	4322.60	3892	3920.40	3947
20×12×5	6808	6923.80	6956	7032.35	7093	7591.15	6842	6963.95	7038
30×10×3	5808	5855.45	5790	5843.20	5989	6219.40	5796	5834.50	5880
30×15×6	11739	11846.60	11649	11846.80	12100	12653.85	11674	11823.65	11902
40×10×2	4848	4906.90	4851	4893.55	4850	5013.95	4793	4806.90	4822
40×15×3	7751	7842.85	7633	7717.50	7764	8019.25	7592	7676.65	7718
50×10×3	8832	8908.75	8933	8998.10	8788	8930.65	8689	8702.10	8717
50×20×4	12968	13092.75	12937	13017.45	12903	13446.35	12855	12993.00	13072
Average	6892.80	6967.71	6921.10	6998.38	7063.50	7388.80	6834.80	6900.46	6950.70

Table 2. Comparisons of T_{avg} and *SD* of HGA, HGA_V, Random+MN3+SO and SHPBIL

Prob- lems	HGA		HGA_V		Random+MN3+SO		SHPBIL	
	T_{avg}	<i>SD</i>	T_{avg}	<i>SD</i>	T_{avg}	<i>SD</i>	T_{avg}	<i>SD</i>
10×5×3	7.20	10.31	18.89	17.94	0.01	95.03	3.40	11.27
10×8×6	75.85	44.31	25.51	34.77	0.13	264.25	24.89	42.36
20×10×3	79.45	18.77	34.58	29.20	0.23	200.31	32.08	17.97
20×12×5	469.17	53.04	179.67	83.29	1.66	223.28	156.61	47.26
30×10×3	299.85	29.43	128.10	50.73	1.05	143.00	107.62	25.70
30×15×6	954.68	89.09	766.40	80.79	16.01	340.50	683.41	59.84
40×10×2	74.81	28.16	74.16	14.65	0.51	94.01	69.85	8.99
40×15×3	362.55	60.45	261.33	41.42	4.81	199.69	215.31	34.50
50×10×3	697.84	40.87	537.64	44.55	4.67	80.72	455.03	8.43
50×20×4	2058.76	76.48	1301.29	61.74	40.65	355.98	1266.78	64.41
Average	508.02	45.09	332.76	45.91	6.97	199.68	301.50	32.07

BEST and *SD* obtained by SHPBIL are much better than those obtained by HGA, HGA_V and Random+MN3+SO. Moreover, the *WORST* values of SHPBIL are smaller than the *AVG* values of the other compared algorithms for almost all the instance. Thus, SHPBIL is an effective algorithm for the m -machine reentrant permutation flow-shop scheduling problem.

5 Conclusion and Future Work

This paper proposed a self-adaptive hybrid population-based incremental learning algorithm (SHPBIL) to solve the m -machine reentrant permutation flow-shop scheduling problem (MRPFSSP). In SHPBIL, the initial population was generated by using several presented methods, the global search was performed through the improved PBIL with adaptive learning rate and mutation scheme, and a local search was guided by the critical-path-based neighborhood. Since the search behavior was enriched as well as global exploration and local exploitation were well balanced, MRPFSSP can be solved effectively. Simulation results and comparisons based on a set of randomly-generated instances showed the effectiveness of SHPBIL. Our future work is to develop some PBIL-based algorithms to deal with re-entrant job-shop scheduling problems.

Acknowledgments. This research was partially supported by National Science Foundation of China (No. 60904081) and 2012 Academic and Technical Leader Candidate Project for Young and Middle-Aged Persons of Yunnan Province (No. 2012HB011).

References

1. Choi, S.W., Kim, Y.D.: Minimizing Makespan on an M-machine Re-entrant Flowshop. *Computers & Operations Research* 35(5), 1684–1696 (2008)
2. Choi, S.W., Kim, Y.D.: Minimizing Makespan on a Two-machine Re-entrant Flowshop. In: *Proceedings of the Fifth Asia Pacific Industrial Engineering and Management Systems Conference*, vol. 31(19), pp. 1–10 (2004)
3. Chen, J.S., Pan, J.C.H., Lin, C.M.: A Hybrid Genetic Algorithm for the Re-entrant Flowshop Scheduling Problem. *Expert Systems with Applications* 34(1), 570–577 (2008)
4. Chen, J.S., Pan, J.C.H., Lin, C.M.: Solving the Reentrant Permutation Flow-shop Scheduling Problem with a Hybrid Genetic Algorithm. *International Journal of Industrial Engineering* 16(1), 23–31 (2009)
5. Baluja, S.: *Population-based Incremental Learning: a Method for Integrating Genetic Search based Function Optimization and Competitive Learning*. Technical Report CMU-CS-94-193. Carnegie Mellon University, Pittsburgh (1994)
6. Salhi, A., Rodriguez, J.A.V., Zhang, Q.F.: An Estimation of Distribution Algorithm with Guided Mutation for a Complex Flow Shop Scheduling Problem. In: *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*, London, UK, pp. 570–576 (2007)

7. Wang, S., Wang, L., Zhou, G., Xu, Y.: An Estimation of Distribution Algorithm for the Flexible Job-shop Scheduling Problem. In: Huang, D.-S., Gan, Y., Gupta, P., Gromiha, M.M. (eds.) ICIC 2011. LNCS (LNAI), vol. 6839, pp. 9–16. Springer, Heidelberg (2012)
8. Pang, H., Hu, K., Hong, Z.: Adaptive PBIL Algorithm and Its Application to Solve Scheduling Problems. In: Proceedings of the 2006 IEEE Conference on Computer Aided Control Systems Design, Munich, Germany, pp. 784–789 (2006)
9. Hong, S.K.: Shape Optimization of Electromagnetic Devices Using Immune Algorithm. *IEEE Transactions on Magnetics* 33(2), 1876–1879 (1997)
10. Baluja, S., Caruana, R.: Removing the Genetics From the Standard Genetic Algorithm. In: Proceeding of the International Conference on Machine Learning, Lake Tahoe, CA, pp. 38–46 (1995)
11. Grabowski, J., Wodecki, M.: A very Fast Tabu Search Algorithm for the Permutation Flow Shop Problem with Makespan Criterion. *Computers & Operations Research* 31(11), 1891–1909 (2004)

A Model of Emotional Intelligent Agent for Cooperative Goal Exploration

Takashi Kuremoto¹, Tetsuya Tsurusaki¹, Kunikazu Kobayashi²,
Shingo Mabu¹, and Masanao Obayashi¹

¹ Graduate School of Science and Engineering, Yamaguchi University,
Tokiwadai 2-16-1, Ube, Yamaguchi, 755-8611, Japan
{wu,mabu,m.obayas}@yamaguchi-u.ac.jp

² School of Information Science & Technology, Aichi Prefectural University,
Ibaragabasama 1522-3, Nagakute, Aichi, 480-1198, Japan
kobayashi@ist.aichi-pu.ac.jp

Abstract. In this study, a novel model of the intelligent agent is proposed by introducing a dynamic emotion model into conventional action selection policy of the reinforcement learning method. Comparing with the conventional Q-learning of reinforcement learning, the proposed method adds two emotional factors in to the state-action value function: “arousal value” factor which affects motivation of action and “pleasure value” factor which influences the probability of action selection. The emotional factors are affected by the other agents when multiple agents exist in the perception area. Computer simulations of pursuit problems of static/dynamic preys were performed and all results showed effectiveness of the proposed method, i.e., faster learning convergence was confirmed comparing with the case of conventional Q-learning method.

Keywords: intelligent agent, emotion model, circumplex model, reinforcement learning, Q-learning.

1 Introduction

Intelligent agent is defined as an autonomous individual in artificial intelligence research. It has the abilities of adaptive learning of behaviors, knowledge extraction, task execution, and so on. As an abstract functional system, intelligent agent has been widely applied to many fields such as autonomous robot, software engineering, network diagnose, etc. In this paper, we limit the intelligent agent to be an internal model of autonomous mobile robots.

In fact, cooperative behaviors are expected when multiple autonomous robots execute some common tasks [1] - [4]. Unknown environment exploration [2], pursuit problem [3], static goal explorations [4] - [7], and simultaneous localization and mapping [8] have been accelerated by the cooperative behaviors of multiple agents. As a direction of autonomous robot development, it is encouraged to introduce biological analogies and social science theories into the design of the architectures

and the function realizations [1]. In [2], Asada et al. realized cooperative behavior acquisition of mobile robots for environment exploration using a reinforcement learning (RL) algorithm [9]. In [3], Kobayashi et al. improved Q-learning, a well-known conventional RL, by adopting attention degree between agents and successfully tackled a pursuit problem shown by computer simulation. Except the concept of value function of state-action given by rewards from the environment, psychology approach using emotional factors are also proposed to define the function of action decision of intelligent agents. In [8], Watada et al. proposed an internal model of mobile robot with emotional factors and heterogeneous artificial neural networks (ANNs) such as self-organizing map (SOM) and multi-layer perceptron (MLP) to raise the exploration efficiency in large scale unknown environments. Ide and Nozawa group proposed a series action rule to drive autonomous mobile robots to navigate an unknown environment perceiving the emotional indexes of other robots [4] [5]. The emotional indexes were two principle dimensions “pleasure” and “arousal” of Russell’s “circumflex model of affect” [10] [11]. In our previous work, we added another psychology factor “curiosity” into Ide and Nozawa’s model to solve the dead-lock problem when multiple agents exploring complex environment with multiple goals [6] [7]. However, these models drive agents to explore the environment continuously but lack of knowledge exploitation. Furthermore, these models were applied to explore static goal(s) without any investigation of dynamic goal (prey) pursuing.

In this paper, we propose a novel internal model of emotional agents by fusing reinforcement learning method and emotional factors. Russell’s circumplex model of affect is also used to decide the action function of agents. For example, when plural agents explore/pursue a goal or a prey (static or dynamic), the distance and the affect of other agents may involve an agent to act different movement. Meanwhile, the state-action-value function of the conventional reinforcement learning (Q value), which expresses the fitness of an action to a state according to the evaluation of environment, is improved to be “state-emotion-action value function” in the proposed model. Different from [4] – [7], the emotional action rules are proposed here originally, and simulation of static and dynamic pursuit problems were executed to show the effectiveness of the proposed model.

2 Q-Learning

Reinforcement learning is a kind of machine learning inspired by the natural mechanism of the brain [9] [12]. Different from the supervised learning and unsupervised learning of machine learning methods, it is a kind of goal-directed learning, and uses the reward from the environment of agent, a learner, to improve the action policy of the learner. Stochastic action policies are usually to be adopted to make the learning select various available actions (behaviors) to explore the unknown states of environments. And kinds of learning algorithms of RL have been proposed to modify the policies to find the optimal solution of a series decisions process. Q-learning algorithm is one of well-known reinforcement learning and it has been widely applied to many fields such as game theory, optimal control, time series forecasting, and multi-agent systems since 1980s.

2.1 Q-Learning Algorithm for Multiple Agents

To an arbitrary state $s_i \in S$ of agent $i \in \{1, 2, \dots, N\}$, an action $a_i \in A \equiv \{a_i^1, a_i^2, \dots, a_i^{|A|}\}$ may be selected from the available action set with a stochastic policy, i.e., by a probability $P(a_i | s_i)$ according to a Boltzmann distribution:

$$P(a_i | s_i) = \frac{\exp(Q(s_i, a_i)/T)}{\sum_{b \in A} \exp(Q(s_i, b)/T)} \quad (1)$$

Where T is a positive parameter (named “temperature”), $Q(s_i, a_i)$ is a state-action value function calculated with its initial value and renewing formula (learning rule).

To modify the policy given by Eq. (1), Q-learning algorithm is given as Algorithm 1 to raise the probability of adaptive action selection.

Algorithm 1

Step 1 Initialize $Q(s_i, a_i)$.

Step 2 Repeat following in each episode.

2.1 Return to the initial state.

2.2 Repeat the following till the final state.

2.2.1 At a observed state $s_i \in S$, select an action $a_i \in A$ according to the policy (Eq.(1)).

2.2.2 Execute the action a_i , observe a reward r and next state s_i' .

2.2.3 Renew $Q(s_i, a_i)$ according to following:

$$Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \alpha[r_i + \gamma \max_a Q(s_i', a_i') - Q(s_i, a_i)] \quad (2)$$

2.2.4 $s_i \leftarrow s_i'$

Where “episode” means one exploration period of the agent, i.e., from the start state to the last state defined by the designer: usually “achieving to goal” or “limited exploring steps”. State s and action a may be in high-dimension vector spaces, whereas Q value $Q(s_i, a_i)$ and reward r_i are scalars.

Algorithm 1 is the conventional Q-learning algorithm [9] except it modifies all agents in a multi-agent system.

3 A Motivation-Action Model

3.1 Russell’s Circumplex Affect Model

Though there are many psychological models to describe the mental world of human being, we use Russell’s circumplex affect model [10] [11] to generate motivation of agent’s action in this study. The model is shown in Fig. 1.

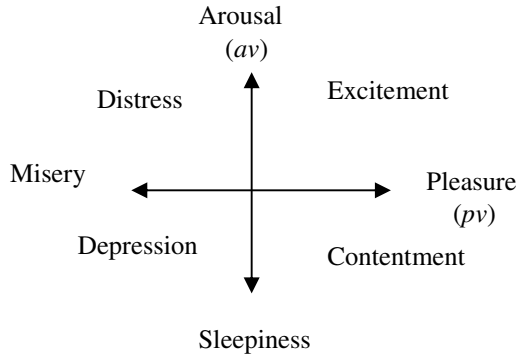


Fig. 1. Russel’s circumplex model of affect [10]

Different from other traditional theory of affect, Russell extracted 8 kinds of factors as shown in Fig. 2 according to 28 stimulus words presented to 36 young people: Arousal, Excitement, Pleasure, Contentment, Sleepiness, Depression, Misery, and Distress. The more interested feature of this model is that all these affect categories are ordered in a 2-D plane with a circular arrangement. So the relationship between them can also be expressed, for example, high values of Arousal and Pleasure may generate a high value of Excitement.

3.2 A Novel Dynamic Emotion Model

Pursuit problem, or hunting game, can be designed in various conditions [3], and for convenience, it is defined as that a static or dynamic goal (prey) needs to be captured by all of plural agents (hunters) simultaneously as a final state here. In Fig. 2, all cases of the final states of a 2-hunter (○) and one prey pursuit problem are shown.

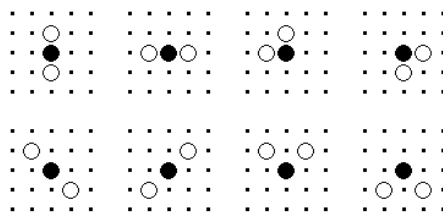
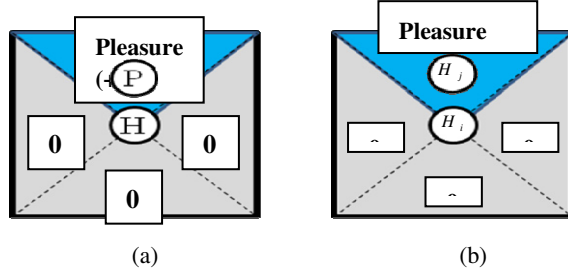


Fig. 2. Final states of a pursuit problem: ○ denotes hunter, ● prey

Here we propose a novel dynamic emotion model to drive hunters to capture a prey.

Definition 1. Emotion of a hunter is limited with one factor: PLEASURE.

Definition 2. AROUSAL provides the multiplier effect to the emotion factor PLEASURE.



H : hunter P : prey Square: vision (perceptible and effect) area of a hunter

Fig. 3. Emotion changes according to the existence in the vision area. Areas with “0” have no change of emotion because no any object is perceived. (a) When a prey P is perceived, $PLEASURE$ of a hunter H $Emo(s_i, a_i) = toPv(s_i, a_i)$ is increased in the direction a_i . Emotion in other areas does not change; (b) When other hunter H_j is perceived, $PLEASURE$ of H_j $Emo(s_j, a_j) = getPv(s_j, a_j)$ is observed and used as $PLEASURE$ of H_i (H_i may be increased or decreased according to $Emo(s_j, a_j)$)

Suppose the vision field of the hunter is a square area divided into 4 directions as shown in Fig. 3. Emotion of the hunter is effected by the object in the vision area. So $PLEASURE$ which means emotion of hunter i can be denoted by a high-dimension vector $Emo(s_i, a_i)$, where $s_i \equiv (a_i, prey \parallel hunter)$ express the state of agent i defined by the vision areas in observal directions (areas) a_i and existance of preys or other hunters, a_i is also the direction of the movement (action) of hunter i , e.g., $|A| = 4, a_i = 1, 2, 3, 4$ for up, down, left right directions, was used in the simulation.

In Fig. 3(a), a case of prey exists in a view area is shown, meanwhile a case of other hunter exists is in Fig. 3(b). Let “prey makes $PLEASURE$ increase”, i.e., $Emo(s_i, a_i) = toPv(s_i, a_i)$, and “emotion of other hunter can be perceived and used” $Emo(s_i, a_i) = getPv(s_i, a_i)$ be emotion rules, then $Emo(s_i, a_i)$ can be calculated as following:

$$Emo(s_i, a_i) = \begin{cases} toPv(s_i, a_i) & \text{if } 0 < d_{prey}(a_i) \leq View, \\ getPv(s_i, a_i) & \text{otherwise} \end{cases}, \quad (3)$$

$$toPv(s_i, a_i) = av_i \cdot pv_i \quad \text{if } 0 < d_{prey}(a_i) \leq View, \quad (4)$$

$$getPv(s_i, a_i) = av_i \cdot pv_j \quad \text{if } 0 < d_{hunter}(a_i) \leq View. \quad (5)$$

Where $d_{prey}(a_i)$ is the distance (Euclidean) between the prey and hunter i in direction a_i and $View$ is a threshold. av_i denotes $AROUSAL$ value of hunter i , pv_i is $PLEASURE$ value changing according to the distance between positions (coordinates) on the vision field (including all directions) and hunter i .

$$av_i \leftarrow \begin{cases} av_i + \Delta av & \text{if } s_i = s_i', \\ av_i - \Delta av & \text{otherwise} \end{cases}, \quad (6)$$

$$pv_i = \begin{cases} P_v \cdot \exp\left(-\left(d_{prey}(a_i)\right)^2 / \sigma^2\right) & \text{if } 0 < d_{prey}(a_i) \leq View \\ -P_v & \text{otherwise} \end{cases}. \quad (7)$$

Where Δav , P_v , σ are positive parameters, s_i and s_i' are the current and next states perceived by the hunter respectively.

Eq. (6) encourages the hunter acts more excited when the state does not change from current to the next time step ($s_i = s_i'$), and reduces *AROUSAL* av_i otherwise.

Eq. (7) controls the value of concerning with the distance between prey and hunter i : more closed more pleasure, and the *PLEASURE* is for all directions.

Summarize the above, hunters act according to the emotional rules as follows.

- (a) A hunter can perceive objects (prey or hunter) in the field of vision and the field is dived in different areas in different directions. The perceive areas are also emotional areas where the hunter's emotion (*PLEASURE*) effects other hunters and is effected by other hunters (Fig. 3; Eq. (3)).
- (b) *AROUSAL* of a hunter changes according to the change of state observed by the hunter (Eq. (6)).
- (c) A hunter becomes to be more pleasure when a prey is perceived in the view area, and the pleasure is ratio to the distance between the hunter and the prey (Eq. (7)).
- (d) Emotion of a hunter is decided by two multiplied factors: *AROUSAL* and *PLEASURE* (Eq. (3)).

3.3 Fusion of Dynamic Emotion Model and Q-Learning

Conventional affect evoked models such as in [4] [5] and [6] [7] have no learning function. Meanwhile, reinforcement learning (RL) usually only uses rewards from external environment [9]. However, in the natural learning process, a higher animal modifies its behavior according to its internal response to the stimulus of the external world: the response of basal ganglia of the brain [12]. So we propose to fuse the dynamic emotion model designed in the last subsection and a conventional RL Q learning described in Section 2.

In Fig. 4, the architecture of emotional intelligent agent is shown. Agent (learner) senses the state of environment s firstly, and then it outputs its action (behavior) a , which may result the change of state. Meanwhile, reward r (or punishment) from the environment is also received by the agent as the evaluation of output (actions).

Instead of $Q(s_i, a_i)$ (Eq. 2), a state-emotion-action value function $M_i(s_i, Emo(s_i, a_i), a_i)$ can be used in the stochastic policy (Eq. (1)):

$$p_i(a_i | s_i, Emo(s_i, a_i), a_i) = \frac{\exp\{M_i(s_i, Emo(s_i, a_i), a_i) / T\}}{\sum_{b \in A} \exp\{M_i(s_i, Emo(s_i, a_i), b) / T\}}, \quad (8)$$

$$M_i(s_i, Emo(s_i, a_i), a_i) = L \cdot Q(s_i, a_i) + (1.0 - L) \cdot Emo(s_i, a_i), \quad (9)$$

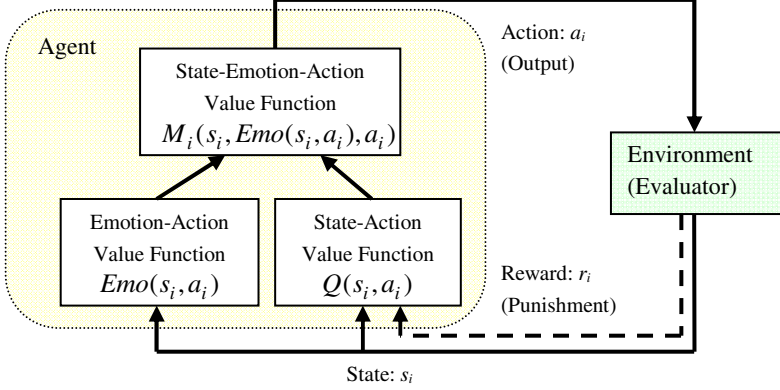


Fig. 4. Architecture of the proposed emotional intelligent agent

where $0 \leq L \leq 1.0$ is a constant to adjust the balance of emotional action and learning exploitation. T is a positive constant called “temperature” and higher T encourage a hunter selects available action more randomly. In our simulation, we used $T \leftarrow T^{Episode}, 0 < T < 1, Episode = 1, 2, \dots, Episode$ to reduce T according to the increase of learning iteration ($Episode$) to obtain better learning convergence.

Notice that the perceive direction “ a_i ” is also used as the movement direction of the hunter, i.e., the action number of the hunter ($a_i \in A$).

The learning algorithm which is a process to modify Eq. (9) is as same as Q learning, i.e., Algorithm 1 described in Section 2.1.

4 Simulation Experiments

To confirm the effectiveness of the proposed model, computer simulations of pursuit problems were executed. In Fig. 5, a case of 2-hunter-1- prey problem with its initial state of the exploration environment is shown. The size of the environment is 17×17 grids. Static/Dynamic simulations were performed which were set as a prey stopped/started at the position (9, 9) at all time, while two hunters start to find and capture the prey from (2, 2) and (14, 14) respectively.

A hunter observes its local environment as a state in 5 dimensions: Wall or Passage in 4 directions (up, down, left, and right) and Prey information in one dimension. For the first 4 dimensions, there are 3 discrete values: near, far, and unperceivable. Whereas Prey dimension has 5 values: up, down, left, right, and unperceivable. So the number of states of a hunter is $4 \times 3 \times 5 = 60$. For the number of action is designed as 4 (up, down, left, right), the number of state-action value function value $Q(s_i, a_i)$ for 2 hunter has $60 \times 4 \times 2 = 480$.

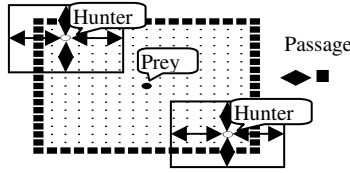


Fig. 5. An exploration environment for 2-hunter-1 -prey problem. Prey stopped or moved to left and right periodically.

Table 1. Common conditions in the simulations

Description	Symbol	Value
Episode limitation	<i>Episode</i>	200 times
Exploration limitation in one episode	step	10,000 steps
Size of the environment	$X \times Y$	17×17 grids
Threshold (depth) of vision field	<i>View</i>	4 grids
Number of hunter	N	2, 3, 4
Number of action	a	4

Table 2. Parameters used in the simulation of 2-hunter-1-static-prey problem

Parameter	Symbol	Q-learning	Proposed model
Learning rate/Damping constant	α / γ	0.9/0.9	0.9/0.9
Temperature (initial value)	T	0.994	0.994
Reward of prey captured by 2/1 hunters	r_1 / r_2	10.0/1.0	10.0/1.0
Reward of one step movement	r_3	-0.1	-0.1
Reward of wall crash	r_4	-1.0	-1.0

Q-learning method and proposed method were compared their learning performance for the pursuit problem in simulation. A part of setting of simulations and parameters are listed in Table 1 and Table 2. A part of special parameters of proposed dynamic emotion model are listed in Table 3. All parameters were obtained by the experimental results: the optimal values were searched. For example, the coefficient of emotion “ L ” used 0.5, which was found by using values from 0.1 to 0.9.

Table 3. Parameters of proposed model used in the simulations

Parameter	Symbol	Value
Coefficient of <i>Emo</i>	L	0.5
Coefficient of <i>PLEASURE</i>	P_v	200.0
Initial value of <i>AROUSAL</i>	A_v	1.0
Modification of <i>AROUSAL</i>	Δav	0.01
Constant of Gaussian function	σ	8.0

4.1 Results of Static/Dynamic Prey Problems

Both Q-learning and the proposed model achieved to the final states when learning process converged in the simulation. The learning convergence of the proposed method (M) was faster and better than Q-learning (Q) as shown as in Fig. 6. The average steps of one episode (from start to final states) during total learning process (200 episodes) given by 10 times simulations of the conventional Q-learning and proposed method including the cases of 3 and 4 hunters were compared in Table 4. Proposed method was confirmed its superiority of learning performance by all simulations.

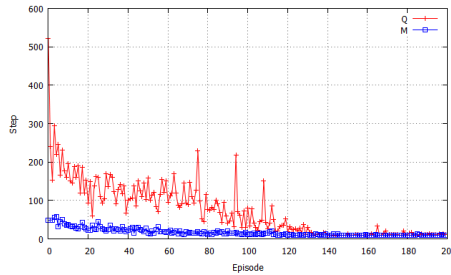


Fig. 6. Comparison of the change of exploration costs (the number of steps from start to final states) during learning process in 2-hunter-1-static-prey problem simulation

Table 4. Average exploration steps during learning in static prey problem simulations (200 episodes, 10 trials)

Number of hunter	Q-learning (static/dynamic prey)	Proposed model (static/dynamic prey)
2	72.9/202.7	17.1/24.4
3	41.1/65.1	15.0/13.3
4	50.7/96.0	18.6/15.3

5 Conclusions

A model of emotional intelligent agent is proposed by introducing a novel dynamic emotion model in to a reinforcement learning algorithm. The original dynamic emotion model used pleasure and arousal factors to affect agents find and capture the static/dynamic prey in the exploration environment cooperatively. The emotion of an agent is formulated and calculated with the distance between itself and the prey. Emotion value of other agent in the view area is used to affect the action selection policy of the learner directly. And these interacted emotion values are adopted into a state-emotion-action value function to improve the conventional Q function of Q-learning.

Simulations of pursuit problems with static and dynamic preys with different number of hunters were executed and all results showed the effectiveness of the

proposed model. The more comparisons between the proposed method and other learning methods for multi-agent systems are expected in the future.

References

1. Cao, Y.U., Fukunaga, A.S., Kahng, A.B.: Cooperative Mobile Robotics: Antecedents and Directions. *Autonomous Robots* 4, 7–27 (1997)
2. Asada, M., Uchibe, E., Hosoda, K.: Cooperative Behavior Acquisition for Mobile Robots in Dynamically Changing Real Worlds via Vision-based Reinforcement Learning and Development. In: *Artificial Intelligence*, pp. 275–292 (1999)
3. Kobayashi, K., Kurano, T., Kuremoto, T., Obayashi, M.: Cooperative Behavior Acquisition in Multi-agent Reinforcement Learning System Using Attention Degree. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) *ICONIP 2012, Part III*. LNCS, vol. 7665, pp. 537–544. Springer, Heidelberg (2012)
4. Sato, S., Nozawa, A., Ide, H.: Characteristics of Behavior of Robots with Emotion Model. *IEEJ Transaction on Electronics, Information and Systems (EIS)* 124(7), 1390–1395 (2004)
5. Kusano, T., Nozawa, A., Ide, H.: Emergent of Burden Sharing of Robots with Emotion Model. *IEEJ Transaction on Electronics, Information and Systems (EIS)* 125(7), 1037–1042 (2005)
6. Kuremoto, T., Obayashi, M., Kobayashi, K., Feng, L.-B.: Autonomic Behaviors of Swarm Robots Driven by Emotion and Curiosity. In: Li, K., Jia, L., Sun, X., Fei, M., Irwin, G.W. (eds.) *LSMS 2010 and ICSEE 2010*. LNCS (LNBI), vol. 6330, pp. 541–547. Springer, Heidelberg (2010)
7. Kuremoto, T., Obayashi, M., Kobayashi, K., Feng, L.-B.: An Improved Internal Model of Autonomous Robot by a Psychological Approach. *Cognitive Computation* 3(4), 501–509 (2011)
8. Watada, S., Obayashi, M., Kuremoto, T., Kobayashi, K.: A New Decision-Making System of an Agent Based on Emotional Models in Multi-Agent System. In: *Proceedings of 18th International Symposium on Artificial Life and Robotics (ARBO 2013)*, pp. 452–455 (2013)
9. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. The MIT Press (1998)
10. Russell, J.A.: A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39(6), 1161–1178 (1980)
11. Larsen, R.J., Diener, E.: Promises and Problems with The Circumplex Model of Emotion. In: Clark, M.S. (ed.) *Rev. Personality and Social Psychology: Emotion*, vol. 13, pp. 25–59 (1992)
12. Doya, K.: Metalearning and Neuromodulation. *Neural Networks* 15(4), 495–506 (2002)

The Cloud Model Based on Grey System Theory and Application on Effectiveness Evaluation

Yanbin Shi¹, Hong Liu², and Jian Sun³

¹ Information Electronic Warfare Department, Aviation University of Air force, Changchun, 130022, China

² Intelligence Institute, Air Force Equipment Research Academy, Beijing, 100085, China

³ Flight Training Base, Air Force Aviation University, Fuxin, Liaoning, China 123100
shiyabin_80@163.com

Abstract. Attention is concentrated on the Cloud model based on grey system theory and the normal grey whitenization weight function used in conversion between qualitative and quantitative uncertainty model. A theoretical treatment of a new model of complex systems' effectiveness evaluation is given. The simulations on different whitenization weight functions have been carried out. Suggestions were made for further study of the Cloud model based on grey system theory and its application on effectiveness evaluation.

Keywords: cloud model, grey system, effectiveness evaluation, whitenization weight function.

1 Introduction

When considered of the complex systems' effectiveness evaluation, because of the internal or external disturbances and the cognitive limitations of the evaluation systems, the information gotten by researchers usually have some uncertainties. Different to the probability statistics and fuzzy mathematics, the grey system theory focuses on solving their difficulty in the "small sample" and "poor information" uncertainties problems. Through generate and develop the little known information to descript the real problems. The evaluating process of the complex system effectiveness not only existence of a large number of fuzzy indexes which have clear connotation but denotation, but also have many "small sample" and "poor information" uncertainties problems[1-4]. Usually, the membership function is used to deal with these problems, but once quantified these indexes by membership degree, they no longer have ambiguous and lost the flexible features on the qualitative evaluation of complex system, and also not accordance with the natural language habits[5].

Cloud is one kind uncertain transition model expressed in natural language between a qualitative concept and its quantitative representation, or the Cloud model is used to achieve the conversion between qualitative and quantitative uncertainty model. This paper use the Cloud model based on grey theory in study of complex system effectiveness evaluation, to solve the uncertainties.

2 Grey Whitenization Weight Function

The grey system is expressed by grey numbers, grey equations, grey matrixes and grey functions, and the grey number is the basis element [6-12]. Due to the limitations of cognitive ability in the research to the complex system, it is difficult to fully recognize the system, but only get some system elements or the range of indicators, these numbers(elements or indicators) are named as grey number, and usually expressed by “ \otimes ”. It can be seen that grey number is a number of sets, the interval of possible values are named whitenization values. The grey number x_i is expressed as $\otimes(x_i)$ and its whitenization value is expressed as $\tilde{\otimes}(x_i)$. According to the uncertainty of information, the role of the grey number’s whitenization value is different, which is expressed by the whitenization weight, and the grey function is named as whitening function and expressed as $f(x)$. So the whitenization weight function descripts different “preferred” degree to the different number in the range.

$$\tilde{\otimes}(x_i) = x_i \cdot f(x_i) \quad (1)$$

Typically, the description of whitenization weight function is straight, when there is more complete information to supplement, according to specific conditions, it can also be curved. But in practice, in order to facilitate the programming and computing, the traditional whitenization weight function is often simplified to linear or triangular whitenization weight function, which has four basic forms used to express the different grey concepts.

3 Cloud and Cloud Model

Cloud is an uncertainty transition model between some quality concepts and its quantity expression expressed by the natural language value, that is the Cloud model is an uncertainty model used to realize the transition between quality and quantity [13-14]. Suppose U is a theory domain expressed by precise value, in the quantity concept A corresponding to U , there is a stable random number $y = \mu_A(x)$ tended to a random element x in the theory domain, named the Membership Degree of concept A , and the distribution of x in the theory domain is called Membership Cloud, shortened for Cloud.

The Cloud is combined by many Cloud droplets. One Cloud droplet may be footy, but the whole shape of the Cloud reflects the important characters of the quantity concepts[15]. Cloud is used to compare to an uncertainty mapping between quantity and quality because this distribution looks like the Cloud in the sky, and it has a clear shape when far looking, but there is no borders in nearly. The digital character of Cloud can be expressed by Ex (Expected value), En (entropy) and He (Hyper entropy).

Expected value Ex is the center value of concept in the theory domain, and it is the representative value of the qualitative concept.

Entropy En is the fuzziness degree of the qualitative concepts, reflects the numerical range which can be accepted by this concept in the theory domain, and embodies the uncertain margin of the qualitative concept. The bigger the entropy is, the bigger numerical range can be accepted by the concept, and the fuzzier are the concepts.

Hyper Entropy He reflects the dispersion of the Cloud droplets. The bigger the Hyper Entropy is, the bigger of its dispersion and the randomness of degree of membership, and so is the Cloud's thickness.

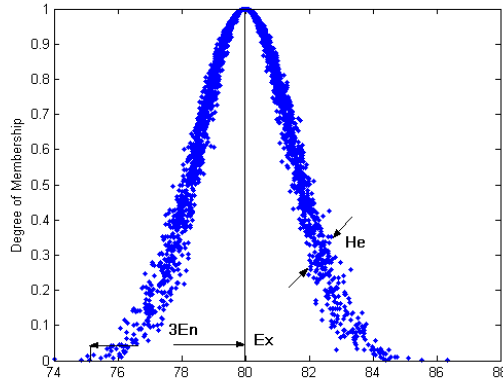


Fig. 1. The normal Cloud and digital characters

So, the fuzziness (the uncertainty of qualitative concept) and randomness (randomness of degree of membership) are completely combined together by three characteristic numbers of the Cloud models, which make up the mapping between quality and quantity as the foundation of knowledge expression.

The normal Cloud is the most basic tool to express the language value, and can be generated by the Cloud's digital character (Ex, En, He), and its mathematic expectation curve (MEC) is:

$$MEC_A(x) = \exp[-(x - Ex)^2 / (2En^2)] \tag{2}$$

The generating algorithm of the normal Cloud is that:

- (1) $x_i = G(Ex, En)e^{i\theta}$, generating a normal random number x_i , whose expected value is Ex and standard deviation is En ;
 - (2) $E'_{ni} = G(En, He)$, generating a normal random number E'_{ni} , whose expectation value is En and standard deviation is He ;
 - (3) compute $\mu_i = \exp[-\frac{(x_i - E_x)^2}{2E'_{ni}}]$, (x_i, μ_i) is a Cloud droplet.
 - (4) Repeat the step 1 to step 4, until the Cloud droplets were generated enough.
- Fig.1 gives out "80" languages which were described by one-dimension Cloud.

The arithmetic of the Cloud is the algorithm changing the base Cloud to the empty Cloud. The empty Cloud means the digital characteristics of some already designated Clouds will carry on a certain operation, Cloud getting the new digital characteristic and constructing. These designated Clouds used for calculating the empty Cloud are called the base Cloud. Because in exceeding the evaluation research of the effectiveness of the complex systems, only use the weighted sum algorithm, so only introduce addition and multiplication of the Cloud here.

If there is a given Cloud in fixed domain $C_1(E_{x1}, E_{n1}, H_{e1})$, $C_2(E_{x2}, E_{n2}, H_{e2})$, the calculated result of arithmetic operation for C_1 and C_2 is $C(E_x, E_n, H_e)$.

Its operation of multiplication can be expressed as:

$$E_x = E_{x1} \times E_{x2}; \quad (3)$$

$$E_n = |E_{x1}E_{x2}| \times \sqrt{\left(\frac{E_{n1}}{E_{x1}}\right)^2 + \left(\frac{E_{n2}}{E_{x2}}\right)^2}; \quad (4)$$

$$H_e = |E_{x1}E_{x2}| \times \sqrt{\left(\frac{H_{e1}}{E_{x1}}\right)^2 + \left(\frac{H_{e2}}{E_{x2}}\right)^2}; \quad (5)$$

Its operation of addition can be expressed as:

$$E_x = E_{x1} + E_{x2}; E_n = \sqrt{E_{n1}^2 + E_{n2}^2}; H_e = \sqrt{H_{e1}^2 + H_{e2}^2} \quad (6)$$

4 The Normal Grey Whitenization Weight Function

Traditional whitenization weight function has been widely applied in various grey decision support systems. But since it only deals with one aspect of uncertainty, processed on the grey concept (imperfect information), so existence of their limitations. In many actualities, the qualitative grey judgment caused by imperfect information and randomness are symbiotic, the grey number usually is a random number within a certain range. While the traditional whitenization weight function is too rigid, and it cannot flexible describe the grey concept.

In recent years, many scholars have begun to discuss the improvement of traditional whitenization weight function, for example in discussion, Dr. Wang Hongli[16] applied the random number generated by Cloud model to replace the traditional whitenization weight function in the grey decision. Then, based on this research, many scholars use the expectation of a number of whitenization weight function generated at random as eventually whitenization weight value, and used in effectiveness evaluation of radar simulation training system, but these preliminary of discussion just considered the standard grey Cloud model. This paper addresses the limitations of traditional whitenization weight function, introduced the Cloud model to the grey whitenization weight function, improved the traditional whitenization weight function, some technologies are used to lay down the single side grey Cloud

model based on non-uniform whitenization weight function[17-19]. In particular, the upcoming whitenization weight function that is made up of Cloud dropletlets with some random curve of non-uniform thickness, with peak grey Cloud model as an example, as shown in Fig.2.

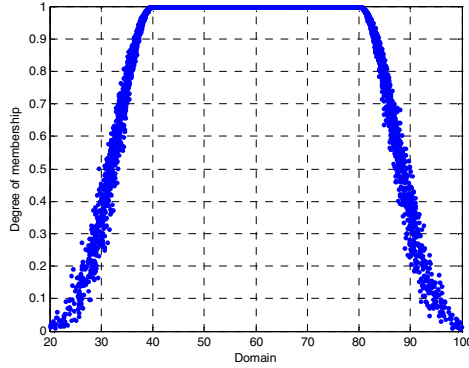


Fig. 2. The normal grey classical whitenization weight function

The normal grey classical whitenization weight function is shown as Fig.2 and equation(7):

$$f(x) = \begin{cases} 0 & x \notin [x_1, x_4] \\ \exp\left[-\frac{(x_i - x_2)^2}{2E_n'^2}\right] & x \in [x_1, x_2] \\ 1 & x \in [x_2, x_3] \\ \exp\left[-\frac{(x_i - x_3)^2}{2E_n'^2}\right] & x \in [x_3, x_4] \end{cases} \quad (7)$$

The normal grey lower limit measure whitenization weight function is shown as Fig.3 (left) and equation(8):

$$f(x) = \begin{cases} 0 & x \notin [0, x_4] \\ 1 & x \in [0, x_3] \\ \exp\left[-\frac{(x - x_3)^2}{2E_n'^2}\right] & x \in [x_3, x_4] \end{cases} \quad (8)$$

The normal grey medium measure whitenization weight function is shown as Fig.1 and equation(9):

$$f(x) = \begin{cases} 0 & x \notin [x_1, x_4] \\ \exp\left[-\frac{(x - x_2)^2}{2E_n'^2}\right] & x \in [x_1, x_4] \end{cases} \quad (9)$$

The upper limit measure whitenization weight function is shown as Fig.3 (right) and equation(10):

$$f(x) = \begin{cases} 0 & x < x_1 \\ \exp\left[-\frac{(x-x_2)^2}{2E_n^2}\right] & x \in [x_1, x_2] \\ 1 & x > x_2 \end{cases} \quad (10)$$

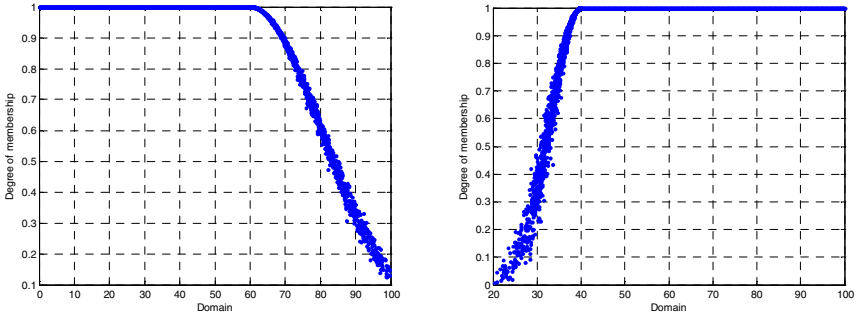


Fig. 3. The normal grey lower and upper limit whitenization weight function

5 The Two-Dimension Normal Grey Whitenization Weight Function

In the complex systems' effectiveness evaluation, many concepts are determined by some related factors, for example, geographic position is determined by longitude and latitude, chromatic color is determined by red, green and blue and their corresponding theory domains are two-dimensional and multi-dimensional theory domain. Along with the thought of the one dimensional Cloud model above, we can extend the Cloud model to two-dimension or multi-dimension, and make it suitable for describing the qualitative concepts of the two-dimensional and multi-dimensional language values.

The two-dimensional Cloud model, which is described by the language value, is an uncertainty transition model between some qualitative concepts and its two-dimensional value. The two-dimensional Cloud concept can be painted as a three-dimensional figure, and Fig.5 is the exterior view of the two-dimensional Cloud which corresponds to the center of the language value. From Fig10, we can see that it looks like a tomb or hill, and the top and foot are very smooth, while the surface of the hillside is rough and protean. It shows that the two-dimensional Cloud's thickness is non-uniform, the hillside is dispersive and the top and foot are more gathered. So, the two-dimensional Cloud is the natural extending of the one dimensional Cloud.

Suppose that the two dimensions of the two-dimensional theory domain are uncorrelated, the two-dimensional normal Cloud can be described by six characteristic numbers $(E_x, E_{nx}, H_x, E_y, E_{ny}, H_y)$. Where, E_x and E_y are expected values, E_{nx} and E_{ny} are entropy, H_x and H_y are hyper entropy. (E_x, E_y) is the central value in the theory domain of the quality concept which is

described by the two-dimensional Cloud, and it's the best to represent the value of two-dimension concept. E_{nx} and E_{ny} are the fuzziness measure of the concept, they reflect the coverage of the language value to the two-dimension value. H_x and H_y are the entropy of E_{nx} and E_{ny} , respectively, and reflect the dispersion degree of Cloud droplets.

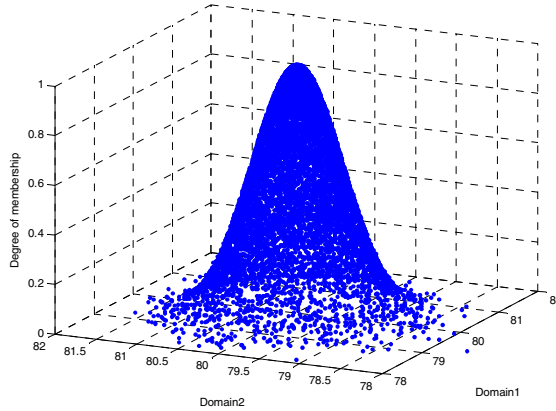


Fig. 4. The two-dimension normal grey whitenization weight function

The algorithm of the two-dimensional positive normal Cloud generator is as follow:

$(x_i, y_i) = G(E_x, E_{nx}, E_y, E_{ny})$, generating a two-dimensional normal random number (x_i, y_i) , whose expected value is (E_x, E_y) and standard deviation is (E_{nx}, E_{ny}) . The specific implementation is that: firstly, generate two one-dimensional standard random numbers t_0 and t_1 , and then compute $x_i = E_{nx} \cdot t_0 + E_x$, $y_i = E_{ny} \cdot t_1 + E_y$, so (x_i, y_i) is the satisfactory two-dimensional random number.

$(E_{nx_i}, E_{ny_i}) = G(E_{nx}, H_x, E_{ny}, H_y)$, generating a two-dimensional standard random number (E_{nx_i}, E_{ny_i}) , whose expected value is (E_{nx}, E_{ny}) and standard deviation is (H_x, H_y) .

$$\text{Compute } \mu_i = \exp \left[-\frac{1}{2} \left[\frac{(x_i - E_x)^2}{E_{nxi}^2} + \frac{(y_i - E_y)^2}{E_{nyi}^2} \right] \right], (x_i, y_i, \mu_i) \text{ is Cloud droplets.}$$

6 Conclusions

Cloud model represents a qualitative concept with three digital characteristics, expected value, entropy and hyper entropy, which integrate the fuzziness and randomness of a linguistic term in a unified way. This paper presents the fundamentals of Cloud theory and its applications in effectiveness evaluation, focusing on the Cloud models and their algorithms. A theoretical treatment of a new model of complex systems' effectiveness evaluation is given. The simulations on different whitenization weight functions have been carried out. The Cloud model based on grey system theory can be used widely in effectiveness evaluation, spatial data mining and knowledge discovery or intelligent decision support system.

References

1. Guang, D.L., Guang, L.L., Zhang, A., Shi, Y.B.: Grey comprehensive relational analysis of fighting efficiency influencing factors of AEW radar. In: 2011 International Conference on Electric and Electronics, pp. 753–760 (2011)
2. Shi, Y.B., Gao, X.J., Zhang, A.: Application of Cloud theory on aviation EW system effectiveness evaluation. *Computer Engineering and Applications* 22, 241–243 (2009)
3. Yang, S.K.: An improved system Cloud grey neural network model. In: Proceedings-2nd International Conference on Genetic and Evolutionary Computing, pp. 158–160 (2008)
4. Wang, G.: Rough set based uncertain knowledge expressing and processing. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS (LNAD), vol. 6743, pp. 11–18. Springer, Heidelberg (2011)
5. Shi, Y., Li, Y., Wang, J.: Effectiveness evaluation about self-screening jamming of assaulting aircraft in complex EM environment. In: Yu, Y., Yu, Z., Zhao, J. (eds.) CSEEE 2011, Part II. CCIS, vol. 159, pp. 360–365. Springer, Heidelberg (2011)
6. Zhao, T.Z., Hu, J.L.: Performance evaluation of parallel file system based on lustre and grey theory. In: Proceedings - 9th International Conference on Grid and Cloud Computing, pp. 118–123 (2010)
7. Chen, Z., Yu, B.F., Hu, W.Y., Luo, X.L., Qin, L.X.: Grey assessment and prediction of the urban heat island effect in city. *Journal of Xi'an Jiaotong University* 38, 985–988 (2004)
8. Chen, Y.: The comprehensive ranking evaluation of flood disaster based on grey-Cloud whitening-weight function. In: Proceedings of 2011 International Conference on Electronic and Mechanical Engineering and Information Technology, vol. 4, pp. 1932–1934 (2011)
9. Zhang, G.X., Zhang, M., Zhong, H.L., Huang, X.Y.: Research of transportation scientific and technological projects evaluation based on improved triangular whitening weight function. In: 2010 International Conference on Logistics Systems and Intelligent Management, vol. 3, pp. 1500–1503 (2010)
10. Liu, J.J., Xiao, M.D., Li, B.J.: An evaluation model based on grey trigonometry whitening weight function for logistic center. In: Proceedings of the 2nd International Conference on Transportation Engineering, vol. 345, pp. 2055–2060 (2009)
11. Zeng, B., Liu, S.F., Cui, J.: Prediction model for interval grey number with known whitening weight function. *Control and Decision* 25, 1815–1820 (2010)
12. Yang, X.W., Wu, S.L., Xu, C.R., Chen, J., Zeng, S.P.: The comparison of two types of whitening weight function. In: 2009 International Conference on Apperceiving Computing and Intelligence Analysis, pp. 69–72 (2009)
13. Wang, J., Xiao, W.J., Zhang, J.L.: An Analysis of an Improved Effectiveness Evaluation based on Cloud Model. *Fire Control & Command Control* 5, 97–99 (2010)
14. Wang, J., Xiao, W.J., Sheng, W.: An Improved Effectiveness Evaluation Method Based on Grey Cloud Model. *Microcomputer Information* 4, 277–278 (2009)
15. Zhang, Y.D., Wu, L.N., Wei, G.: Improved ant colony algorithm based on membership Cloud models. *Computer Engineering and Applications* 27, 11–14 (2009)
16. Wang, H.L.: Grey Cloud model and its application in intelligent decision support system supporting complex decision. In: International Colloquium on Computing, Communication, Control, and Management, vol. 1, pp. 542–546 (2008)
17. Liu, J.J., Wang, W., Cheng, L.: The grey number evaluation of road net based on trigonometry whitening weight (TWW) function. *System Engineering Theory and Practice* 27, 156–159+176 (2007)
18. Xia, X.T., Wang, Z.Y.: Expanded uncertainty model based on whitened weighting function. In: Proceedings of the Second International Symposium on Instrumentation Science and Technology, vol. 3, pp. 935–938 (2002)

Finding the Optimal Bus-Routes Based on Data Mining Method

Yong Wang, Yang Liu, Cheng-zhi Zhang, and Zhi-ping Li

College of Information Science and Engineering,
Guangxi University for Nationalities Nanning 530006, China
wangygxnn@sina.com

Abstract. It is a very difficult for a passenger to find out an optimal bus-route to travel in a metropolis. A novel approach of search the optimal bus-routes based on data mining technology is presented in this paper. In the approach, we first use the Floyd algorithm to search and to get the candidate set of the optimal bus-routes, and then use data mining method to mine the optimal bus-lines which is hidden in the candidate set. Experiment were conducted on a real bus-line data set, and the experimental results shows that the approach is effective and feasible in search for the optimal bus-routes.

Keywords: Optimal bus-route, Data mining, The support of scheme, Optimal index.

1 Introduction

City public-bus traffic system can provide an economical and convenient transportation service for most of citizens. Public-bus has already become the first choice for most local residents when they need to travel from one place to another. Hence, how to improve the efficiency of bus-transportation has become a real problem for the city-project experts and also has been a focus for many scholars[1-9]. For example, Su[3] presented a method to solve the problem of how to get the shortest path of public-bus network based on the improved Dijkstra algorithm. Wang[4] put forward an algorithm about the optimal routes of the public traffic system by using the direct matrix and the minimal transfer matrix. Dai[5] analyzed the reliability of transfer time of public traffic system. Su Xiao[6] devised a relational inquires solution about bus routes. Li[7] presented a public bus transfer automatic text message platform. Chen[8] proposed a shortest-path algorithm which is fitted for the bus inquires system, et al. These approaches which were presented by the above scholars can be summed up as follows. (a)That these methods is based on Dijkstra algorithm. (b)That these methods only discuss the case of need transferring at most two times, or only discuss the problems of the shortest-routes, and did not discussed the case that there are many optimal schemes of bus-lines from one bus-stop to another. However, if we live in a big city, such as in Shanghai, Beijing of China, we will meet with the case that we usually need transferring more than two times to go to our destination from one place to another. So in the following, we will discuss the case that there are

many optimal schemes of bus-lines and need transfer more than one times from one bus-stop to another.

In this paper, we have proposed an approach of finding the optimal bus-routes based on data mining method, using the strategy of combining Floyd algorithm with data mining technology. In the paper, we first use Floyd algorithm to search the candidate set of the optimal lines, and then make use of data mining method to mining the optimal bus-routes which is hidden in the candidate set. The rest of this paper is organized as follows. Section 2 gives our approach. Section 3 gives a real case study. We give our conclusions in Section 4.

2 Finding the Optimal Bus-Routes Based on Data Mining

2.1 Main Theory and Method

Let $G = \langle V, E \rangle$ be a graph of the public-bus network, in which, $V = \{v_1, \dots, v_m\}$ is the bus-stop set of G , $E = \{(v_i, v_j), i, j = 1, \dots, m\}$ is all of the section lines in G , v_i refers to a single bus-stop in G , $e_i = E(v_i, v_j)$ denote the weight or the length of every (v_i, v_j) respectively, and $W = \{e_1, \dots, e_n\}$, in which, m is the total number of the bus stops and n is the sum of the bus lines. Let $l_s = v_{s1}v_{s2} \dots v_{sk_s}$ be a bus route, and suppose l_s is unique, i.e. $v_{si} \neq v_{sj}$ ($i \neq j$). In which, we require, for the l_s , that there only exist under two cases, one case is that v_{s1} is the starting bus stop and v_{st} is the end bus stop, and we describe the length of l_s as $E(v_{s1}, v_{s2}) + \dots + E(v_{s_{k_s-1}}, v_{s_{k_s}})$, where $v_{si} \neq v_{sj}$ ($i \neq j$). If there exists a bus-line l_s through the v_i bus stop and the v_j bus stop, then we call that v_i can reach to v_j directly. Otherwise, we call that v_i can not reach to v_j directly. That is to say, if a passenger wants to get to the v_j bus-stop by bus from the v_i bus stop, he had to change bus to another.

For the sake of convenience, throughout this paper, we use the marks as following. Suppose v_a is a bus-stop, then we denote

$$L_a = \{\text{bus routes of passing the } v_a \text{ bus-stop}\},$$

$$V_{La} = \{v_i | \exists l_i \in L_a, v_i \text{ is one of a bus-stop of } l_i\}.$$

We mark v_s and v_e as the start-bus-stop and the end-bus-stop respectively, and denote

$$L_{se} = \{\text{bus routes of passing } v_s \text{ and } v_e\}.$$

We first investigate whether there is a direct bus-line passing v_s and v_e .

If $L_{se} \neq \phi$, then it means that there exists a direct bus-line through the v_s bus-stop and the v_e bus-stop at least. If $L_{se} = \phi$, then it means that there isn't any straightforward bus-line between the v_s bus-stop and the v_e bus-stop. This time, if we want to

get to the v_e bus-stop from the v_s bus-stop by bus, we had to change from one bus-line to another. Next step, we will discuss how to find out all of the transfer-schemes which can get to the v_e bus-stop from the v_s bus-stop.

If there isn't any direct bus-line between the v_s bus-stop and the v_e bus-stop, then we will investigate how to find out all of the transfer-schemes from the v_s bus-stop to the v_e bus-stop.

We first give the definition of 'join' operation as follows.

Definition1. Let $A_i = \{a_{i1}, \dots, a_{ik_i}\}, i = 1, \dots, w$, then the 'join' operation is defined as

$$\text{join}(A_1, \dots, A_w) = \{a_{1j_1} \dots a_{wj_w} \mid a_{ij_i} \in A_i, i = 1, \dots, w\}.$$

In which, element $a_{1j_1} a_{2j_2} \dots a_{wj_w}$ is an ordered arrangement.

For example, if $A_1 = \{v_s\}, A_2 = \{l_1, l_2\}, A_3 = \{l_3, l_4, l_5\}, A_4 = \{v_e\}$, then

$$\text{join}(A_1, A_2, A_3, A_4) = \{v_s l_1 l_3 v_e, v_s l_1 l_4 v_e, v_s l_1 l_5 v_e, v_s l_2 l_3 v_e, v_s l_2 l_4 v_e, v_s l_2 l_5 v_e\},$$

where every element is a sequence.

Following, we will use the Definition 1 to search for all the bus-lines by only one-transfer.

First, a one-transfer bus-route can be depicted as " v_s (start-stop) $\rightarrow \dots \rightarrow v_p$ (passing-stop) $\rightarrow \dots \rightarrow v_e$ (end-stop)", where, $v_s, v_p \in l_s, v_p, v_e \in l_e, l_s$ is one of the bus-lines passing v_s and v_p, l_e is one of the bus-lines passing v_p and v_e , and $l_s \neq l_e$.

We denote

$L_s = \{\text{bus routes of passing } v_s\}, V_{L_s} = \{v_i \mid \exists l_i \in L_s, v_i \text{ is one of a bus-stop of } l_i\},$
 $L_e = \{\text{bus routes of passing } v_e\}, V_{L_e} = \{v_i \mid \exists l_i \in L_e, v_i \text{ is one of a bus-stop of } l_i\},$
 $V_z = V_{L_s} \cap V_{L_e}$. This time, if $v_i \in V_z$, then v_i is one of the transfer-bus-stops to go to the v_e bus-stop from the v_s bus-stop. Let $\mid V_z \mid = z$.

If $v_p \in V_z$, then $\exists l_{sp} \in L_s, \text{ s.t. } v_s, v_p \in l_{sp}, \text{ and } \exists l_{pe} \in L_e, \text{ s.t. } v_p, v_e \in l_{pe}$. This time, if they take bus according to the schematic line " $v_s \rightarrow \dots \rightarrow v_p \rightarrow \dots \rightarrow v_e$ ", then they can go to the v_e bus-stop from the v_s bus-stop by only one transfer. In which, $v_s \rightarrow \dots \rightarrow v_p$ is a section of the line l_{sp} , and $v_p \rightarrow \dots \rightarrow v_e$ is a section of the line l_{pe} .

This time, $\text{join}(L_{sp}, L_{pe}) = \{\text{all transfer-schemes between } v_s \text{ and } v_e\}$. We can use a graph to depict those transfer-schemes as follows in the Fig.1.

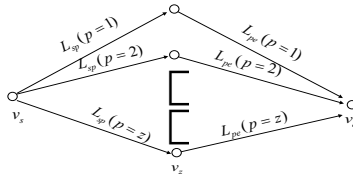


Fig. 1. One transfer-schemes between v_s and v_e

Second, a n - transfer bus-route can be expressed as “ v_s (in l_1 , start-stop) $\rightarrow \dots \rightarrow v_{12}$ (1^{th} transfer-stop,in l_2) $\rightarrow 1 \dots \rightarrow v_{23}$ (2^{th} transfer-stop,in l_3) $\rightarrow \dots \rightarrow l_n \rightarrow \dots \rightarrow v_e$ (end-stop in l_n)”, where $v_{ij} = l_i \cap l_j$ is a transfer-stop. We can use a graph to depict these transfer-bus schemes as below in Fig.2.

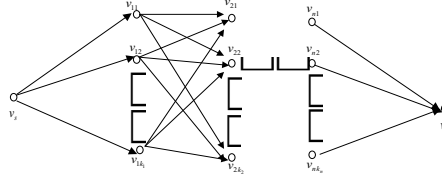


Fig. 2. Multi-transfer-schemes between v_s and v_e

In the above graph, we mark v_{1j_1} as the first transfer-bus-stop, v_{nj_n} as the n^{th} transfer-bus-stop, $j_1 = 1, \dots, k_1, j_n = 1, \dots, k_n$, and $V_{zi} = \{v_{ij} \mid j = 1, \dots, k_i\}, i = 1, \dots, n$. Let

$$L_{s,1j_1} = \{l \mid \exists v \in V_{z1}, l \text{ is a bus-line through } v_s \text{ and } v \text{ directly}\},$$

$$L_{j_i,(i+1)j_{i+1}} = \{l \mid \exists v_{ij_i} \in V_{zi}, v_{(i+1)j_{i+1}} \in V_{z(i+1)}, l \text{ is a bus-line passing } v_{ij_i} \text{ and } v_{(i+1)j_{i+1}} \text{ directly}\},$$

$$L_{nj_n,e} = \{l \mid \exists v \in V_{zn}, l \text{ is a bus-line passing } v \text{ and } v_e \text{ directly}\}.$$

This time, if a person wants to get to the v_e bus-stop from the v_s bus-stop by bus, then his taking-bus line can be described as

$$v_s \rightarrow L_{s,1j_1} \rightarrow v_{z1} \rightarrow L_{1j_1,2j_2} \rightarrow v_{z2} \rightarrow \dots \rightarrow v_{zn} \rightarrow L_{nj_n,e} \rightarrow v_e.$$

Following, we will use ‘join’ operation to get all of the taking-bus schemes. That is to say, we get the set of the taking-bus schemes through ‘join’ operation as follows

$$\text{join}(L_{s,1j_1}, \dots, L_{nj_n,e}) = \{v_{1j_1} v_{2j_2} \dots v_{nj_n} \mid v_{ij_i} \in V_{zi}, i = 1, \dots, n\} \triangleq D_{sn}.$$

We denoted this set as D_{sn} , then D_{sn} is a sequential data set. If the number n is big enough, then D_{sn} will be a massive data set. So we will use data mining approach to excavate the optimal taking-bus lines hidden in the D_{sn} following.

Mining frequent patterns is one of the main tasks for data mining. The sequence pattern mining is to excavate the frequent pattern in a series data set. We first give the conception of the support of a transfer scheme.

Definition 2. Let $S_* = \{l \mid l \text{ is the shortest line between } v_s \text{ and } v_e\}$, and L_{se} be a transfer scheme between v_s and v_e . We call $\max_{l \in S_*} |L_{se} \cap l|$ to be the matching degree of L_{se} . We also call $\max_{l \in S_*} |L_{se} \cap l|$ as the support of L_{se} .

For example, if $S_* = \{l_1, l_2\}, l_1 : v_s \rightarrow v_1 \rightarrow v_2 \rightarrow v_e, l_2 : v_s \rightarrow v_3 \rightarrow v_6 \rightarrow v_e$, and L_{se} is $v_s \rightarrow v_1 \rightarrow v_3 \rightarrow v_4 \rightarrow v_5 \rightarrow v_e$, then $\max_{l \in S_*} |L_{se} \cap l| = 3$.

Next step, we use Floyd algorithm to search for the set of the shortest routes. Suppose $S_* = \{P_k \mid k = 1, \dots, w\}$ is the set of the shortest routes found by Floyd algorithm. If $w = 1$, then it means that there exists a single shortest route from v_s to v_e . We suppose $w \geq 2$, and mark p_k to be the number of bus-stops in the shortest route P_k , where $P_k \in S_*$. We denote $|P_k| = p_k, k = 1, \dots, w$.

We first transform every element in D_{sn} into a sequence data. For example, if $v_{1j_1} \dots v_{nj_n} \in D_{sn}$, then we transform $v_{1j_1} \dots v_{nj_n}$ into a data segment $(v_{1j_1}, \dots, v_{nj_n})$. Let $D = (\text{segment}1, \dots, \text{segment}n)$, where n is the number of transfer schemes. This time, the sequence data D is composed by n data segments and every data segment in D is starting from v_s and ending at v_e . Denote $\max_{P_k \in S_*} |P_k \cap l_i| = s_i$, where l_i is the segment i , and the initial value of s_i is 0, $i = 1, \dots, n$.

2.2 Main Mining Approach

In the beginning, let $p_1 = |P_1|$ be the initial width of sliding window. If the data which is fallen into the sliding window is a subset of P_1 , then let p_1 be the width of sliding window. If $p_1 > s_i$, then let $S_i = p_1$, and scanning the next segment. If the data which is fallen into the sliding window is not a subset of P_1 , then let $p_1 - 1$ be the width of sliding window. Let this segment be continuously scanned until the width of sliding window being 1. This time, let $s_i = 1$, and let the next segment be scanned continuously. Since v_s and v_e both are included in every segment. Therefore, while the width of sliding window is 1, the data which is fallen in the sliding window must include v_s (the start-bus-stop) or v_e (the end-bus-stop). But v_s and v_e both are included in P_1 . So the smallest support of every transfer scheme is not small than 1.

Let $p_k = |P_k|$ be the width of sliding window ($k = 2, \dots, w$). If $k < w$, then scanning the data set D continuously according to the procedure mentioned above until $k = w$. Finally, we get all the maximum matching degree $s_i, i = 1, \dots, n$.

The main mining procedure(or steps) can be described as follows.

Input $D, n, P_k, p_k = |P_k|$.

Output s_i .

Step1. While $i > n$, stop running and output s_i .

If $p_k = 1$, then $S_i = 1$, and turn to **Step3**.

Else, scan the segment i with a sliding window of whose initial width is p_k .

If the data fallen into the sliding window is a subset of P_k , then let $S_i = p_k$, and turn to **Step3**.

Else, turn to **Step2**.

Step2. Let $L_k = L_k - 1$, turn to **Step1**.

Step3. Let $i = i + 1, p_k = P_k \setminus i$, turn to **Step1**.

Step4. If $k > w$, then stop running and output result.

Else $k = k + 1$, turn to **Step1**.

Following, we denote the distance of the transfer-scheme l as d_l , the number of bus-stops in the transfer-scheme l as N_l , and the transfer-times of the transfer-scheme l as T_l . We need to calculate d_l, N_l and T_l . For example, if $l = v_1 v_2 \cdots v_n$ is a transfer-scheme, then $N_l = n$, in which, v_1, v_2, \dots, v_n are all the bus-stops in l . We denote the distance between v_i and v_{i+1} as $E(v_i, v_{i+1})$, where, v_i and v_{i+1} are two adjacent bus stops in $l, i = 1, \dots, n - 1$. This time, $d_l = E(v_1, v_2) + \cdots + E(v_{n-1}, v_n)$.

In order to use data mining technology to mining the optimal bus-routes, we first give the conception of the optimal index of scheme.

Definition 3. We call $T_l(d_l + N_l) / S(l)$ as the **optimal index of the scheme l** . In which, $l = v_1 v_2 \cdots v_n, d_l$ is the distance of the transfer-scheme l, N_l is the number of bus-stops in the transfer-scheme l, T_l is the transfer-times of the transfer-scheme l , and $S(l)$ is the support of scheme l .

We denote $O_l = T_l(d_l + N_l) / S(l)$. This time, if O_l is the smallest one in all of the candidate set, then the scheme l will be the optimal schemes.

3 Real Case Study

Example 1. If there is a public-bus-line network such as in the following Fig.3.

Line 1	Stop C $\xrightarrow{5}$ Stop K $\xrightarrow{4}$ Stop B $\xrightarrow{2}$ Stop F $\xrightarrow{4}$ Stop A
Line 2	Stop A $\xrightarrow{5}$ Stop J $\xrightarrow{3}$ Stop E $\xrightarrow{3}$ Stop I $\xrightarrow{5}$ Stop D
Line 3	Stop F $\xrightarrow{3}$ Stop K $\xrightarrow{5}$ Stop C $\xrightarrow{2}$ Stop H $\xrightarrow{1}$ Stop I
Line 4	Stop I $\xrightarrow{5}$ Stop D $\xrightarrow{2}$ Stop H $\xrightarrow{2}$ Stop C $\xrightarrow{5}$ Stop K
Line 5	Stop K $\xrightarrow{3}$ Stop H $\xrightarrow{1}$ Stop I $\xrightarrow{4}$ Stop J $\xrightarrow{2}$ Stop F

Fig. 3. A real case of public-bus line-net

It is obvious that there is not a direct bus-line between the start-bus-stop E and the end bus-stop K in above Fig.3.

We first use Floyd algorithm to get the candidate set of bus-routes. In this real case, we get 90 two-transfer-schemes and six one-transfer-schemes which are listed in the Fig.4.

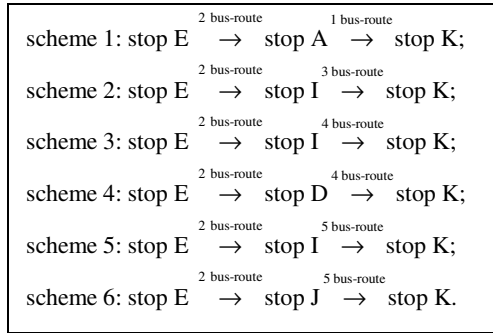


Fig. 4. The results of one-transfer-schemes

Secondly, we use data mining method to mining the optimal bus-routes which are hidden in the Fig.4. We use Definition 3 to get the following Table 1.

Table 1. The results of need one times transfer

scheme	transfer times	length of scheme	number of stop	support of scheme	optimal index
scheme1	1	18	6	1	24
scheme2	1	11	5	3	16/3
scheme3	1	17	6	2	23/2
scheme4	1	17	6	2	23/2
scheme5	1	9	4	4	13/4
scheme6	1	11	5	3	16/3

From the Table 1, we obtain that $O_{\text{scheme 5}} = 13/4 = 3.25$ is the smallest one, so the scheme5 is the optimal transfer-scheme among these six schemes, and the scheme is “stop E →(No.2 bus)→ bus stop I →(No. 5 bus)→ bus stop K”.

From this simple case, we can find if we conduct our experiment on a large and complex public-bus-line network, we will get a large data set of one-transfer-schemes, and a massive data set of multi-transfer-schemes. So it is necessary for us to find out which transfer-scheme is the best one among these transfer-schemes.

4 Conclusions

In this paper, we propose an approach of search the optimal transfer-schemes by using Floyd algorithm combined with date mining method. We conduct an experimental simulating study on a real case of bus-line. The experimental results shows that this approach is feasible and effective in finding the optimal transfer-schemes. Next step, we will try to study how to use this method to search for the dynamic optimal routes from a dynamic bus-line set.

Acknowledgments. This work were supported by the key programs of institution of higher learning of Guangxi of China under Grant No.201202ZD032, the Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, the Natural Science Foundation of Guangxi of China under Grant No.0832084, and the Natural Science Foundation of China under Grant No. 61074185.

References

1. Carr, S.C.: Southeast Florida Regional Travel Characteristics Study. Prepared for Florida Department of Transportation (2000)
2. Stern, R.: Passenger Transfer System Review. Synthesis of Transit Practice 19, Transportation Research Board, National Research Council, National Academy Press, Washington D.C. (1996)
3. Su, A.H., Shi, F.Z.: Optimal route choice of public traffic network based on shortest path searching. *Journal of Engineering Graphics* (4), 55–59 (2005) (in Chinese)
4. Wang, L., Li, W.Q.: Best-routing algorithm for public transportation systems. *Journal of Southeast University (Natural Science Edition)* 2(3-4), 264–267 (2004) (in Chinese)
5. Dai, S., Jiang, H.P., et al.: Study on transfer reliability of integrated public transport. *Journal of Highway and Transportation Research and Development* 24(9), 124–135 (2007) (in Chinese)
6. Su, X., Zeng, Z.W.: Inquiring system of city's bus changing based on related web site. *Computer Engineering and Design* 27(3), 519–521 (2006) (in Chinese)
7. Li, G., Lin, Y.H., Guo, J.G.: Public traffic transfer inquiry system based on the note platform. *Journal of Fujian Agriculture and Forestry University* 37(4), 440–443 (2008) (in Chinese)
8. Chen, X.F., et al.: Shortest path algorithm analysis and its application to bus route query. *Journal of Engineering Graphics* (3), 20–24 (2001) (in Chinese)
9. Wang, J.H.: Study on Transfer System of Urban Passenger Transportation and it's Evaluation. Southwest Jiaotong University Doctor Degree Dissertation (2003) (in Chinese)

Analysis and Application of Data Mining in CRM System of Some Mobile Branch

Song-Tao Lou¹, Ji-Rui Li¹, and Wen-Bin Zheng²

¹Information Engineer Department, Henan Vocational and Technical Institute, Zhengzhou, China

²Computer research section, Armed Police Officers School in Shijiazhuang, Shijiazhuang, China

Abstract. Data mining techniques in customer relationship management (CRM) has a very important role in the process, it is an important means of gaining and maintaining customer information, and improving customer value. This paper expounds the concept of customer relationship management and the process of data mining, analysis business value of data mining in customer relationship management, make clearly the importance of data mining in customer relationship management, and apply it to customer relationship management system of some mobile subsidiary company.

Keywords: Customer relationship management (CRM) Data Mining Mobile Company.

1 Customer Relationship Management (CRM)

Customer relationship management, also referred to as “customer relationship management”. In practice, Customer is translated customer said the meaning of more extensive, it includes the consumers purchase or are buying and “potential customers” who are not buying but may produce purchase behavior in the future, it refers to the more accurate[2].

“Take customer satisfaction as the center” is the core of the customer relationship management. On the basis of it, three basic concepts are derived naturally, the concept of customer value, the concept of relationship value and information value concept. interaction of three value concept, as shown in Fig. 1.

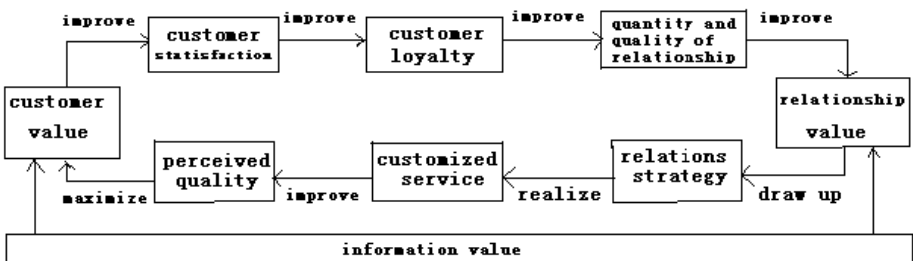


Fig. 1. Derivative concept of CRM

(1) The basis of customer relationship management is to create customer value for clients.

The purpose of customer relationship management is to achieve the balance of between the maximization of customer value and the maximization of the profits. the purpose of any enterprise about making customer relationship management is to create more value for customers, which realizes “win-win” of the customer and enterprise. Adhere to the customer as the center, creating value for customers is the cornerstone of any customer relationship management strategies must have.

(2) To create relationship value for companies is the result of the customer relationship management.

Enterprise is a profit organization, Enterprise's ultimate objective is to realize the maximization of enterprise value, therefore, in building customer relationships, relationship value must be considered. So-called relationship value, namely, to establish and maintain a particular customer relationship that can create new value for the enterprise[4]. Logically, the total value of enterprise should be equal to the sum of all the past, present and future of customer relationship value. Enterprises should focus on high customer relationship value, because it can create the higher profits. And for those customer relationship who value is low, do not have training prospects and may even bring negative effect, enterprise should decisively terminated.

(3) Customer relationship management approach is to make full use of information value.

Information technology is the key factor in customer relationship management. Customer relationship management approach is to make full use of the information value, make it into customer value and relationship value. The emergence of information technology that lets the enterprise to efficiently analyze customer data, gives full play to the role of the information value, incentives and shares customer knowledge, according to the different customer preferences and characteristics to provide the corresponding services, so as to improve customer value. At the same time, information technology can help companies identify with different value of customer relationship, according to different customer relationship adopt different strategies, so as to realize the balance between customer value maximization and profit maximization.

2 Data Mining

2.1 Definitions

Data Mining, that is a nontrivial process from a large amount of Data to obtain valid, novel, potentially useful and ultimately understandable patterns. From the generalized point of data mining, it is stored the "dig" interesting knowledge process a large amount of data in a database, or data warehouse, or other information. Transactions found by data mining are generally shown on the form below: 1, Concepts; 2, Rules; 3, Regularities; 4, Patterns; 5, Constraints; 6, Visualizations[1]. These transactions can be directly provided to decision makers, to assist in the decision-making process; Also can be provided to the domain expert, to correct a system of knowledge formed on the expert; Also can be used to storage mechanism of application System as a new business, such as Expert System, Rule Base and so on.

2.2 The Process of Data Mining

The basic processes and the main steps of data mining is shown an Fig 2[3].

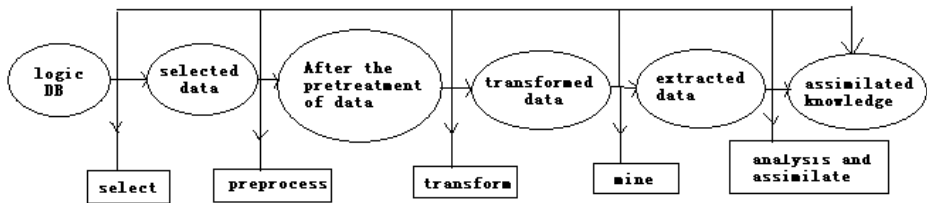


Fig. 2. The basic processes and the main steps of data mining

(1) to transform business problems for data mining problems

Clearly define the business problem, identify that the purpose of data mining is an important step in data mining. the end of the structure in digging is unpredictable, but should have foreseen to explore the question, it is blindness to data mining and data mining, will not succeed.

(2) data preparation

Data preparation include: select data-----extract the set of target data in a large database and data warehouse; preprocessing data-----it is data reprocessing, includes to check the integrity of the data and the consistency of the data, to the noise, to fill the missing fields, and delete the invalid data, etc. To create a truly suitable analysis model of the mining algorithm is the key to the success of data mining. Data preparation is the first step in data mining, also is the more important step. Data preparation is good or bad that will affect the efficiency and accuracy of data mining, and the effectiveness of finally model.

(3) data mining

According to the characteristics of data and the types of data functions, corresponding algorithm would be chosen, for data mining at data sets which is purification and transformation.

(4) results analysis

Interpreting and evaluating the results of data mining, and converting it into knowledge which user understands.

(5) knowledge assimilation

The knowledge gained from analysis should be integrated into the organizational structure of the business information system.

3 CRM System Analysis of a Mobile

The mobile branch employs about 3000, with nearly 50% employees have university degree or above. The main operations are: international and domestic long distance communications, mobile communications, data communications, wireless paging, Internet and electronic commerce and all kinds of telecom value-added business.

3.1 Design Goal of CRM System

In order to provide better services to customers, improve customer loyalty, and reduce the loss of customers, the company put forward to using the data mining model to deepen customer relationship management. Specific targets of customer relationship management system is:

- (1) adopting a small data set of details of the company's current customers and customer has to move as a research sample data sets, analysis client group, in order to provide personalized custom service for valuable customers;
- (2) using the method of data mining to identify customers is not satisfied to company, communication with them, and improve itself services, do everything possible to retain customers, avoid the loss of customers especially big customers;
- (3) to as much as possible reduce the malice owe customer, but at the same time to avoid conflict with customers especially those big customers, requiring the client without loss, reduce the malicious customers owe fee.
- (4) analysis various factors lead to customers to move and understand its importance, thus improve the customer retention rate.

3.2 Function Analysis of CRM System

The system function structure is shown in Fig 3.

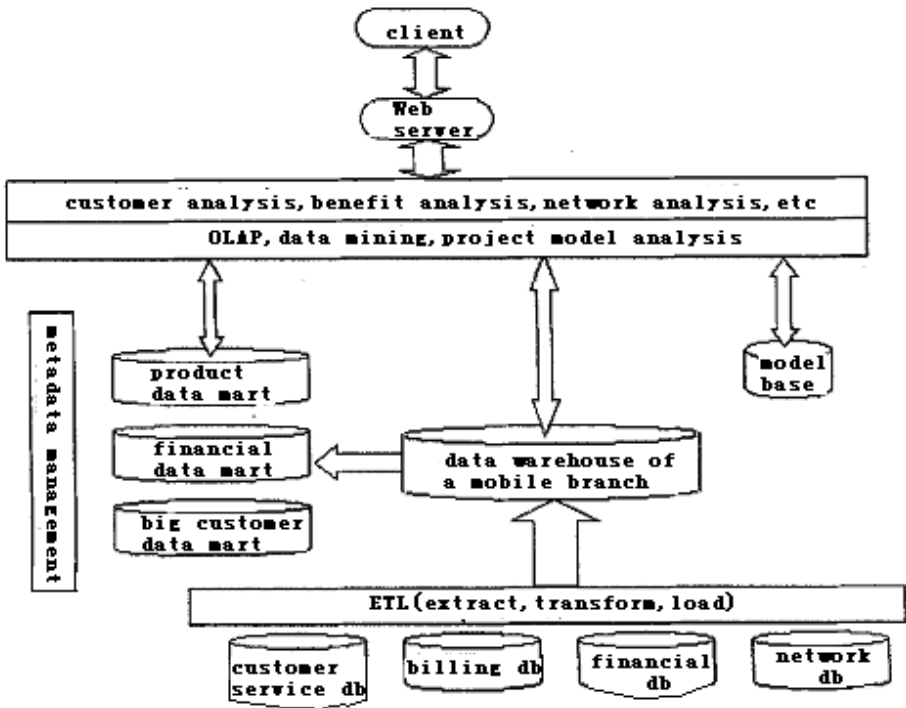


Fig. 3. Functional diagram of CRM

The company customer service center, billing information center, finance department and network management center has its own set of information system, with a lot of business data and dispersion of customer data, but every information system database is not compatible, this is due to historical reasons. Although the branch has a large amount of data, but it is difficult to across departments to get the data, let alone to comprehensive analysis of these data. In order to better regulate all kinds of data, reduce redundancy, the company decided to set up data warehouse, make the data for the theme, thus better able to analyze these data.

The system on the basis of the data warehouse, using data mining technology and the model provided by the system to make the customer development analysis, benefit analysis, the call feature analysis, analysis of marketing management and network optimization. The division of these features stems from the perspective of enterprise operating division, basically covers the enterprise internal data requirements. Using model library at the same time also can analyze customer behavior, respectively is: client group analysis, customer churn analysis, and customer fraud analysis.

3.3 Data Mining Analysis of CRM System

(1) customer groups analysis

Customer group division is the basis of the understanding and optimization of company operation situation. On the basis of reasonable classification to the customer, may adopt different strategies for different customer groups, reasonably guide for the consumption behavior of customer.

Customers can be divided by using clustering and classification mining method. Clustering method is learning without the guidance, which divide on the basis of specified rules not conducted; Classification method is learning of guiding, in the specified rules, to judge and predict the customer may belong to a group. Because the company has specified the rules of customers division in accordance with their own classification requirements, so using the classification method. As shown in Fig 4, by prediction finding the middle customer of 19% may soon become a big customer, so the company can take some preferential policies to guide them into the company's big customer.

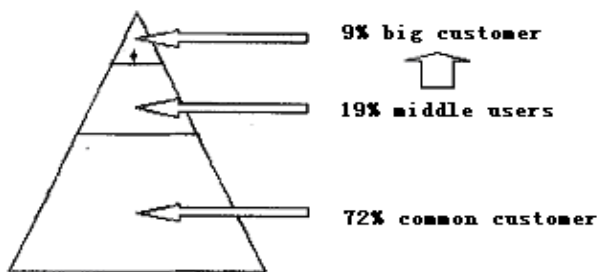


Fig. 4. Classification of customer base

(2) customer churn analysis

Because competition and change of the communications market, the possibility of customer to rival grew there. Customer churn is a very serious problem for telecom company. As is known to all, cost to keep an old customer than to win a new customer is much lower.

Through analysis call behavior of customer and data mining, "anomalies" shows before the customer losses which can be found, the company improve the service quality, resolve complaints, price reductions and other measures to retain customers. As shown in Fig 5, through the model analysis, found that complaints of a and b gold customers increased rapidly in recent times, and the call number is falling fast. loss of gold customers is very big to the company's loss, special attention should be.

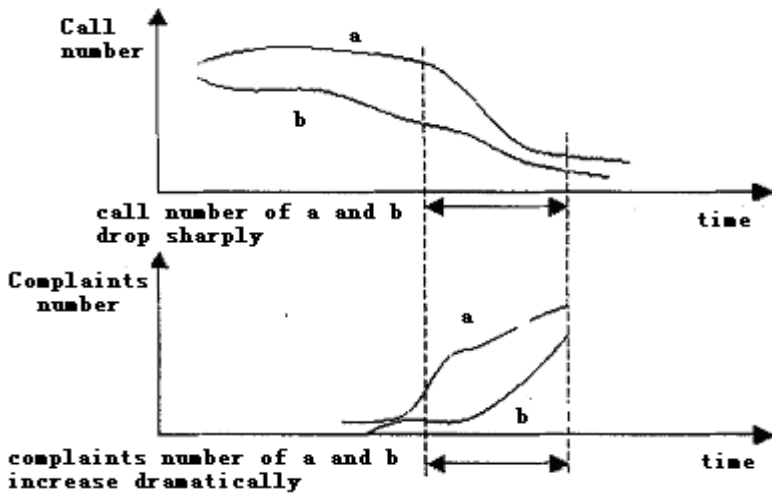


Fig. 5. customer churn analysis

(3) customer fraud analysis

Customer fraud refers to the customer that do not comply related agreement with the business, fail to perform the obligations, and make businesses suffering some losses, merchants were affected by the fraud. Customer fraud is very common, bringing the loss of the enterprise is very big. For the telecommunications industry, the owe phenomenon is more prominent, which not only make the enterprise back heavy burden, also make the enterprise is difficult to establish a good marketing strategy.

Fraud can be divided into: malicious fraud, revenge fraud and non malicious fraud. Fraud analysis is mainly based on the client's calling behavior and other related factors. As shown in Fig 6, from the data analysis, it reflects call charges of the user a has increased dramatically in the short term, referencing other factors, the user a had free records, considering its vocational characteristics, it is not to engage in sales or related occupations, so it can be list in suspicious users. If complaint the user a in the near future also rose sharply, it will explain the possibility of retaliation fraud.

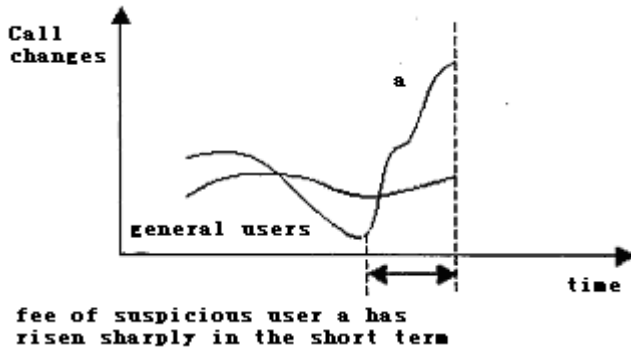


Fig. 6. customer fraud analysis

Companies should give high attention for these phenomena, a reasonable analysis model which can find maximally real malicious users, at the same time avoid too sensitive, so as not to make unpleasant confrontation with big and reputable customers.

4 Summarize

Customer relationship management is not only a kind of marketing strategy, but is a business revolution geared to all walks of life and involving all the business processes of company. How to statistic and analysis market and customer information by data mining technology, find market opportunities, identify the target customer group and the marketing mix, scientifically formulate marketing and product strategy are more and more concern of the modern enterprise. China's accession to the WTO, make Chinese enterprises will speed up the learning of management experience and the pace of information construction, and strive to reach international standard. customer relationship management on information means is the best way to help companies achieve this goal. At present, our customer relations management is in the phase of enlightenment. Therefore, studying the core theory is important significance to the development of customer relationship management and enterprise informational level on our country.

References

1. Feng, Y.: Application of Data Mining in CRM System, pp. 6–11. Northeast DianLi University, Ji Lin (2007)
2. Customer Relationship Management (CRM) And Data Mining Technology (2012), <http://www.doc88.com/p-703879380530.html>
3. Zhang, Y.: Expert System Research of Roll Better Educated Based on Artificial Intelligence. Hebei University of Science and Technology, 16–19 (2011)

4. Li, Z.: Research of CRM System Based on Data Mining, pp. 131–137. Hebei University (2010)
5. Tan, Z.: Huge Amounts of Data Mining Algorithms Research Based on Rough Set. Chongqing University of Posts and Telecommunications, 34–37 (2004)
6. Zheng, Y., Zhang, L.Z.: Mining Interesting Locations And Travel Sequences from GPS Trajectories. In: Proceedings of the 20th International Conferences on World Wild Web, pp. 791–800 (2009)

Line Clipping Algorithm of Affine Transformation for Polygon

Wenjun Huang

College of Information Science and Engineering Guangxi University for Nationalities,
Nanning, 530006

Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis,
Nanning, 530006

hwjart@126.com

Abstract. Applying affine transformations to polygon (concave or convex) and line segment, the algorithm of this paper obtains the points of intersection of the polygon and the line segment. Having tested the line segment with a bounding box which includes the polygon, the algorithm applies the affine transformation to the polygon and the line segment; in the polygon, the algorithm finds the edges which intersect the straight line. The algorithm applies another affine transformation to the edges, and obtains the points of intersection of the polygon and the line segment. Finally, The algorithm applies the reverse affine transformations to the points, and obtains the intersection points of the original polygon and the original line segment.

Keywords: Clipping, Affine transformation, Algorithm, Polygon.

1 Introduction

Line clipping is a basic operation of computer graphics. The Line clipping of convex polygon has fully been discussed[1, 2, 3]. The study for the line clipping of arbitrary polygon has gotten obvious progress[4,5,6,7,8,9]. All the methods in the above papers have to solve equations to find the intersections of the polygon and the line segment, but the method of the paper[9] is novel. It reduce the difficulty of line clipping by affine transformations, but it still solve equations to find the intersections of polygon and line segment.

Our paper will makes a new method. The method does not solve equations. The method applies two affine transformations to arbitrary polygon and line segment to get the intersection points.

2 Theorem

Apply the affine transformation to the oblique line, and the line may become horizontal or vertical.

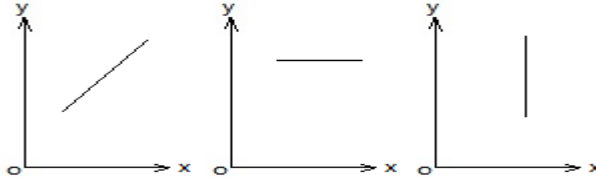


Fig. 1. Having gotten the affine transformation, the oblique line become the horizontal or vertical line

Proof: Make the oblique line as follows: $((x_0, y_0), (x_1, y_1))$.

Take k as the slope of the line segment.

(1) make the affine transformation T_y as follows:

$$x' = x, \quad y' = y - kx.$$

Apply the affine transformation to the line segment L , and the line segment L become the line LL as follows:

$$((x_0, -k \cdot x_0 + y_0), (x_1, -k \cdot x_1 + y_1)).$$

make the comparison as follows:

$$\textcircled{1} x_1 - x_0 \neq 0;$$

$$\textcircled{2} (-k \cdot x_0 + y_0) - (-k \cdot x_1 + y_1)$$

$$= y_0 - y_1 + k \cdot x_1 - k \cdot x_0$$

$$= y_0 - y_1 + k(x_1 - x_0)$$

$$= y_0 - y_1 + (x_1 - x_0) \cdot (y_1 - y_0) / (x_1 - x_0)$$

$$= 0$$

So the line LL is horizontal.

(2) make the affine transformation T_x as follows:

$$y' = y, \quad x' = x - (y/k).$$

Apply the affine transformation to the line segment L , and the line segment L become the line LL as follows:

$$(-(y_0/k) + x_0, y_0), \quad (-(y_1/k) + x_1, y_1).$$

make the comparison as follows:

$$\textcircled{1} y_1 - y_0 \neq 0;$$

$$\textcircled{2} (-(y_0/k) + x_0) - (-(y_1/k) + x_1)$$

$$= y_1/k - y_0/k + (x_0 - x_1)$$

$$= (y_1 - y_0)/k + (x_0 - x_1)$$

$$= (y_1 - y_0)(x_1 - x_0) / (y_1 - y_0) + (x_0 - x_1)$$

$$= (x_1 - x_0) + (x_0 - x_1)$$

$$= 0$$

So the line LL is vertical.

(3) Apply the affine transformation T_y (T_x) to a vertical (horizontal) line, the line is still vertical (horizontal);

3 The Algorithm Based on the Theorem

In the section, we only study the line segment in the situation of $|dy| > |dx|$. The algorithm may be applied to the situation of $|dy| \leq |dx|$ directly.

0) Make a line segment L and a polygon W .

- 1) Move the line segment and the polygon, and set one endpoint of the line segment on the origin of the rectangular coordinate system.
- 2) Apply the affine transformation T_x to the line L and the polygon W . Having gotten the affine transformation T_x , the line L overlap the axis Y .
- 3) Search the edges of the polygon; find the edges which intersect the axis Y ; apply the affine transformation T_y to the edges.
- 4) Having gotten the affine transformation T_y , the edges are vertical to axis Y . In the set of the intersection points of the edges and the axis Y , find the valid intersection points, and apply the inverse transformations of the affine transformation T_x and the move to the valid intersection points, and obtain the intersection points of the line segment and the polygon.

See Figure. 2.

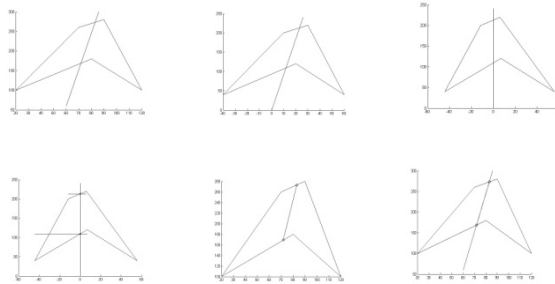


Fig. 2. (a).The situation of $|dy| > |dx|$

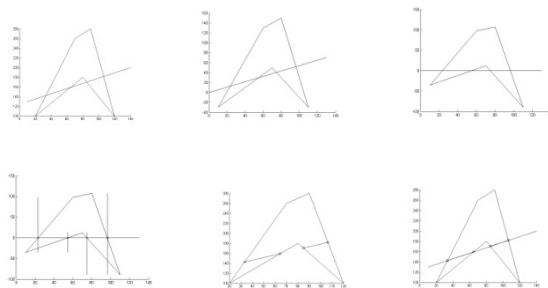


Fig. 2. (b).The situation of $|dy| \leq |dx|$

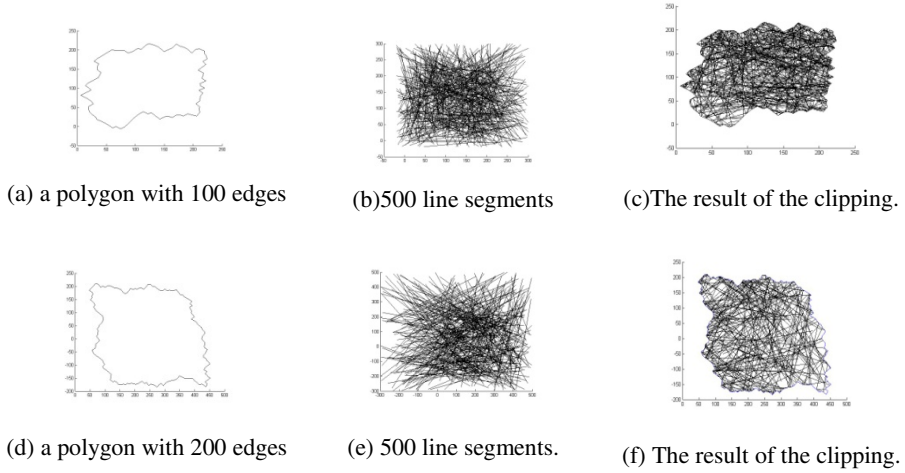


Fig. 3. Two examples of the experiment

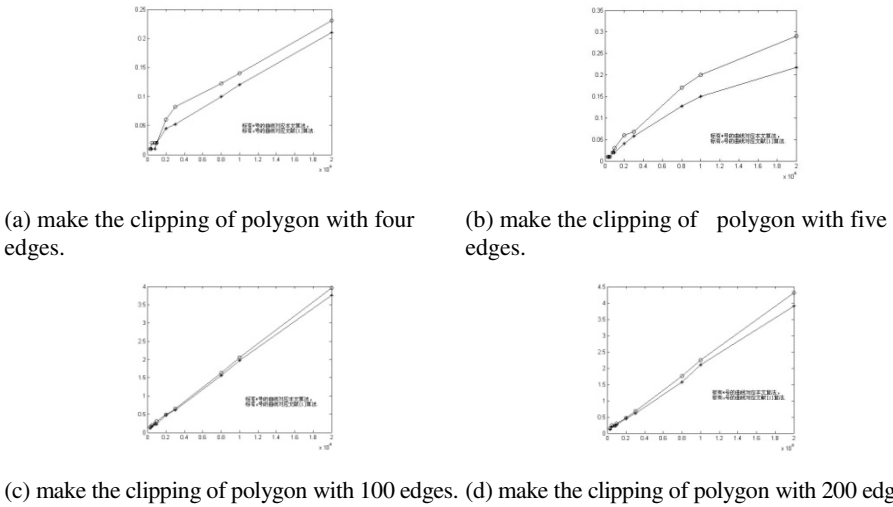


Fig. 4. The curves of the time of running the two algorithms. The curve with the symbol "*" corresponding to our algorithm. The curve with the symbol "o" corresponding to the algorithm of paper[9].

Table 1. The comparison of our algorithm and the algorithm of paper[9] for polygon with 4 edges

	The number of line segments									
	150	250	400	500	1000	1500	4000	5000	10000	
The time of the algorithm of the paper[9]	0.01	0.02	0.02	0.02	0.0601	0.0826	0.1227	0.1402	0.2303	
The time of our algorithm	0.01	0.01	0.01	0.02	0.0450	0.0526	0.100	0.1202	0.2103	
The time of the algorithm of the paper[9]										
/ The time of our algorithm	1	2	2	1	1.34	1.57	1.23	1.17	1.1	

Table 2. The comparison of our algorithm and the algorithm of paper[9] for polygon with 5 edges

The number of line segments	150	250	400	500	1000	1500	4000	5000	10000
The time of the algorithm of the paper[9]	0.01	0.01	0.02	0.03	0.0601	0.0676	0.1703	0.2003	0.2904
The time of our algorithm	0.01	0.01	0.02	0.02	0.0401	0.0576	0.1277	0.1502	0.2178
The time of the algorithm of the paper[9]									
/ The time of our algorithm	1	1	1	1.5	1.5	1.17	1.33	1.33	1.33

Table 3. The comparison of our algorithm and the algorithm of paper[9] for polygon with 100 edges

The number of line segments	300	500	800	1000	2000	3000	8000	10000	20000
The time of the algorithm of the paper[9]	0.1477	0.1878	0.2454	0.3130	0.4882	0.6434	1.6273	2.0505	3.9582
The time of our algorithm	0.1227	0.1702	0.2253	0.2278	0.4782	0.6109	1.5572	1.9653	3.7604
The time of the algorithm of the paper[9]									
/ The time of our algorithm	1.2	1.10	1.09	1.37	1.02	1.05	1.05	1.04	1.05

Table 4. The comparison of our algorithm and the algorithm of paper[9] for polygon with 200 edges

The number of line segments	300	500	800	1000	2000	3000	8000	10000	20000
The time of the algorithm of the paper[9]	0.1577	0.2428	0.2303	0.3004	0.4732	0.6785	1.7525	2.2407	4.3037
The time of our algorithm	0.1252	0.1928	0.2354	0.2629	0.4481	0.6084	1.5723	2.0949	3.8981
The time of the algorithm of the paper[9]									
/ The time of our algorithm	1.23	1.26	0.99	1.14	1.06	1.12	1.11	1.07	1.10

4 Experiment for Comparison

We make the program of our algorithm in MATLAB language, and compare to the algorithm of the paper[9]. The parameters of the computer for the experiment are as follows: Intel Pentium, 1.60GHz CPU, 191 MHz, 252MB memory. We make four polygons and nine sets of line segments by the function `randint()`. One polygon deal with the nine sets of line segments. There are two examples of the experiment in Fig. 3. We get the time of running the algorithms by the function `cputime()`, and we put the time in table 1 ~table 4. Fig.4 show the comparisons of the algorithm in paper[9] and our algorithm.

The data of the tables show that the speed of our algorithm is higher than the algorithm of paper[9].

5 Conclusion

The algorithm of this paper for polygon to clip line segments is valid. The algorithm does not need to deal with the equations to find the intersections. The algorithm finds the intersections by affine transformations. The speed of running the algorithm is higher than some algorithms.

References

1. Skala, V.: $O(\lg N)$ line clipping algorithm in E2. *Computers & Graphics* 18(4), 517–527 (1994)
2. Skala, V.: A New Approach To Line And Line Segment Clipping in Homogeneous Coordinates. *Visual Computer* 21(11), 905–914 (2005)
3. Tang, J., Zhang, Q., Sun, H.: A Clipping Algorithm Based on Cross Product Method for Convex Polygon Window. *Journal of Northeast Heavy Machinery Institute* 19(1), 23–25 (1995)
4. Han, J., Ge, Y., Zhang, D.: An Algorithm for Vector Graphic Based a polygon Window. *Journal of Taiyuan University of Technology* 36(2), 160–163 (2005)
5. Ren, H.: Line Clipping Algorithm Against a Polygon Based on Point Region Distributing. *Science Technology and Engineering* 9(16), 4833–4835 (2009)
6. Lu, G., Xing, S., Peng, Q.: An Efficient Algorithm of Line Clipping Against Polygonal Window Based on the Vertex Encoding. *Chinese Journal of Computers* 25(9), 987–993 (2002)
7. Li, W.: A Scan Strip Based Algorithm for Line Clipping against Arbitrary Polygons. *Journal of Engineering Graphics* (2), 35–40 (2005)
8. Li, J., Wang, W., Wu, E.: Line Clipping Against a Polygon Based on Convex Decomposition. *Journal of Computer Aided Design & Computer Graphics* 19(4), 425–429 (2007)
9. Huang, Y.Q., Liu, Y.K.: An Algorithm for the Clipping Against a Polygon Based on Shearing Transformation. *Computer Graphics Forum* 21(4), 683–688 (2002)
10. Huang, W.: The Line Clipping Algorithm Basing on Affine Transformation. *Intelligent Information Management* 2(6), 35 (2010)

Evaluating Community Detection Using a Bi-objective Optimization

Nesrine Ben Yahia^{1,2}, Narjès Bellamine Ben Saoud^{1,3}, and Henda Ben Ghezala^{1,2}

¹ Laboratoire RIADI

² Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Tunisia

³ Institut Supérieur d'Informatique, Université Tunis El Manar, Tunisia
{Nesrine.benyahia,Narjes.Bellamine}@ensi.rnu.tn

Abstract. Community detection consists on a partitioning networks technique into clusters (communities) with weak coupling (external connectivity) and high cohesion (internal connectivity). In order to measure the performance of the clustering, the network modularity is largely used, a metric that presents the cohesion and the coupling of communities. In this paper, a global and bi-objective function is proposed to evaluate community detection. This function combines modularity (based on structure and edges weights) and the inter-classes inertia (based on nodes weights). Then, we rely on a computational optimization technique i.e. Particle Swarm Optimization to maximize this bi-objective quality. Finally, a case study evaluates the proposed solution and illustrates practical uses.

Keywords: Community Detection, Modularity, inertia, Particle Swarm Optimization.

1 Introduction

Community detection consists on a networks clustering technique (like computer networks and social networks) into clusters (communities) with weak coupling and high cohesion [1]. Many existing techniques of community detection, like [2], [3] and [4], are structure based techniques where clustering only focuses on topological structure of the network i.e. nodes connectivity however it does not consider the nodes attributes so clusters could be often heterogeneous. Other existing techniques of community detection, like [5] and [6], are attribute based techniques where clustering only focuses on edges weights (similarity between attributes nodes of each edge) but it does not consider their connectivity. Finally, there are structural/attribute based techniques, like [7] and [8], where clustering focuses on both attributes nodes (edges weights) and their connectivity so each partition contains cohesively joined and homogeneous nodes.

In order to evaluate the clustering performance, [1] introduced modularity metric as a measure of the quality of a particular clustering of a network that considers communities coupling (external relations among communities) and cohesion (internal relations within communities). Modularity maximization is one of the most popular

community detection methods [1] and it is used in many works such as [9], [10], [11] and [12].

As modularity optimization is typically NP-hard optimization [13], different NP-hard optimization techniques are used to maximize it. For example [1] and [10] relied on greedy heuristic technique. Others authors relied on Meta heuristic techniques such as the use of genetic algorithm in [14], [15] and [16], ant colony algorithm in [17], [18], [19] and [20] or simulated annealing in [21]. But, a comparative study, done in [22], confirms that, Particle Swarm Optimization (PSO) which consists on a computational technique for NP-hard optimization [23], is faster than the last cited techniques and give best results for optimization.

So, in order to detect strongly connected and homogenous communities with high modularity and high density, we propose in this paper, to combine the network structure, edges weights and nodes weights into a global and bi-objective function. Then, to maximize this function, we rely on Particle Swarm Optimization (PSO) which represents a computational optimization technique.

The outline of this paper is as follows. In section two, we illustrate the crucial concepts involved in community detection. In section three, we propose our bi-objective function to evaluate community detection. Finally, section four presents a case study where we will experiment the proposed solution.

2 Community Detection

In a graph, community represents a dense group of nodes that are more cohesively connected to each other than to the nodes in other communities [24]. Therefore, community consists on a group of nodes which are densely related to each other with better internal connectivity than external connectivity.

2.1 Community Detection Approaches

Approaches of community detection fall into three classes: agglomerative, divisive and modularity optimization [25]:

- Agglomerative approaches, used for example in [26], are based on bottom-up processes that focus on the addition of edges to the network. They start with an empty network (n nodes with no edges) and consider the vertex pairs with highest similarity, calculate distances between nodes pairs, and then add edges to the network.
- Divisive approaches, used for example in [1], are based on top-down processes that focus on the removal of edges from the network. They start with a network of connected nodes and search the least similar connected nodes pairs and then remove the edges between them. By repeating this removal, the network is divided into smaller and smaller communities.
- Finally, modularity maximization approaches, used for example in [9], [10], [11] and [12], are based on optimization techniques to find the best clustering that

maximize the network modularity. This last represents the difference between the fraction of edges in the considered network and expected value of the same quantity in a network with the same community division but random connections between the vertices.

2.2 Community Detection Evaluation: The Modularity

To evaluate the performance of community detection techniques, there are different metrics that are proposed and the common one is the modularity. Suppose we have a network that contains n vertices, and let the number of edges between vertices i and j be A_{ij} , which will normally be 0 or 1, so the quantities A_{ij} are the elements of the so-called adjacency matrix. At the same time, the expected number of edges between vertices i and j if edges are placed at random is $k_i * k_j / 2m$, where k_i and k_j represent the degrees of the vertices and m is the total number of edges in the network. Thus the modularity can be written as follow:

$$Q = \frac{1}{2m} \sum_{i,j}^n \left(A_{ij} - \frac{(k_i * k_j)}{2m} \right) \delta(C_i, C_j) \quad (1)$$

This formula introduced in [1], is given for finding a good division of a network into two parts. As many networks contain more than two communities, [1] proposed a standard approach to this problem that extends the first method to find good divisions of networks into larger numbers of parts. So, the second one is repeated division into two: it uses the first method to divide the network into two parts, then divide those parts, and so forth.

[10] relied on the same formula modularity and confirms that even in the case of weighted networks (weighted networks are networks that have weights on their links) A_{ij} represents the weight of the edge between i and j , k_i is the sum of the weights of the edges attached to vertex i .

Then, as modularity optimization is typically NP-hard optimization [13], different NP-hard optimization techniques are used to maximize this measure. For example [1] and [10] relied on greedy heuristic technique. Others authors relied on Meta heuristic techniques such as the use of genetic algorithm in [14, 15, 16], ant colony algorithm in [17 18 19 20] and simulated annealing in [21].

3 Proposal of a Global Function of Community Detection Evaluation

3.1 Global Quality Formula

In our proposed bi-objective function, we have two parts. One is related to the network structure and edges weights and another is related to nodes weights. For the first part, we reuse the modularity of Newman [1] that considers the edges weights as it is mentioned in [10] for weighted graphs. Then, for the second part, we propose to reuse the notion of inter-classes and intra-classes inertia [27].

Inertia permits to measure the dispersion of the cloud. According to Huygens's formula (2) Inertia can be subdivided in two components: intra-classes inertia and inter-classes inertia.

$$I_{\text{total}} = I_{\text{inter}} + I_{\text{intra}} \quad (2)$$

The intra-class inertia, given by (3), measures the degree of homogeneity between objects belonging to the same class. It calculates their distances from the point representing the profile of the class (the gravity center of the class), p_i and I_i represent respectively the weight and the inertia of each class C_i and $d(x,y)$ is generally the Euclidian distance between x and y .

$$I_{\text{intra}} = \sum_i p_i I_i = \sum_i p_i \sum_{x_j \in C_i} \frac{d^2(x_j, g_i)}{|C_i|} \quad (3)$$

The inter-class inertia, given by (4), measures the degree of heterogeneity between classes. It calculates the distances between the points representing different profiles classes of the partition.

$$I_{\text{inter}} = \sum_i p_i d^2(g_i, g) \quad (4)$$

Therefore, a low value of the inter-classes inertia reflects homogeneity within classes and a high value of inter-class inertia reflects heterogeneity between classes. For example, to obtain a good classification, ward [28] proposed to minimize the intra-classes inertia (or, dually, maximize the inter-classes inertia).

In this paper we introduce a bi-objective function as a weighted combination of Newman modularity and variance explicated by inter-classes inertia (α is a weighting factor where $0 < \alpha < 1$):

$$Q = \alpha \frac{1}{2m} \sum_{i,j}^n \left(A_{ij} - \frac{(k_i * k_j)}{2m} \right) \delta(i, j) + (1 - \alpha) \frac{I_{\text{inter}}}{I_{\text{total}}} \quad (5)$$

3.2 Bi-objective Function Optimization with PSO

To get best clustering, our aim here is to maximize the proposed bi-objective function given by (5), so we have to maximize simultaneously the network modularity and the variance explicated by the inter-classes inertia.

As modularity optimization is typically NP-hard optimization [13], we rely on Particle Swarm Optimization (PSO) which consists on a computational technique for NP-hard optimization [23] as it is introduced in [29].

PSO is introduced in [30] as a new evolutionary Meta heuristic technique. We choose this technique thanks to its simple implementation, its fast convergence and its efficiency [31].

In PSO, the swarm contains particles, candidate solutions to the optimization problem. So each particle proposes and evaluates a solution. Each particle i , for each

iteration t , is characterized by its position $x_i(t)$ and its velocity $v_i(t)$. Each particle updates its velocity thanks to its own experience and the experiences of all swarm particles so they accelerate towards positions with optimized evaluation or fitness function. Consequently, each particle is accelerated towards two kinds of position: personal best position achieved until the current iteration (P_{best_i}) and global best position obtained by all swarm particles (G_{best}) [32].

There are different PSO versions. In this paper, we choose one of a basic version proposed in [33] as it focuses on particles cooperation rather than competition. This PSO version, with constriction factor, is characterized by its convergence speed. In this case, the update of position $x_i(t+1)$ and velocity $v_i(t+1)$ at the iteration $t+1$ for particle i are calculated using the position and velocity at the iteration t , P_{best_i} and G_{best} as shown in the following equations:

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (6)$$

$$v_i(t+1) = K * v_i(t) + c_1 * r_1 * (P_{best_i} - x_i(t+1)) + c_2 * r_2 * (G_{best} - x_i(t+1)) \quad (7)$$

Where K : the constriction factor given by

$$K = \frac{2}{|2 - c - \sqrt{c^2 - 4c}|} \quad (8)$$

- With $c=c_1+c_2$ and $c>4$.
- r_1 and r_2 : are random numbers between 0 and 1.
- c_1 and c_2 : are self and swarm confidence factors.

The evolution procedure of PSO Algorithm can be summed up in the following pseudo code:

- First of all, the problem dimension are fixed (PSO parameters), the positions and velocities initializations are defined.
- The fitness of each particle is calculated.
- Position and velocity of each particle are updated using (6) and (7).
- If the maximum iteration number is reached, the final solutions are given.
- Otherwise, we loop to 2 and so on.

In order to maximize the proposed function, in this paper, each particle in the swarm proposes threshold that clusters the organization into two groups (one contains the nodes where weights exceed the threshold and the other contains the rest of nodes). Therefore, for the present optimization problem, the proposed bi-objective function represents the fitness function to maximize where its value is between 0 and 1 and higher value represents better partition.

The proposed algorithm is illustrated by fig.1. To initialize our PSO algorithm parameters, we propose the following values:

- Initial swarm population contains 10 particles, the number of iterations is 1000, $c_1=2.8$ and $c_2=1.3$ basing on [34].

- Initial positions and velocities of particles are determined by using random values in $[W_{\min}, W_{\max}]$ where W_{\min} and W_{\max} represent respectively the minimal and maximal values of nodes weights (highest and lowest similarities).
- For each particle x_i , initial fitness value is initialized to the bi-objective function corresponding to the threshold proposed by x_i using (5).
- For each particle, local best $P_{\text{best}i}$ is initialized to its current fitness value.
- Global best value of the position G_{best} is initialized to the best value of all $P_{\text{best}i}$.

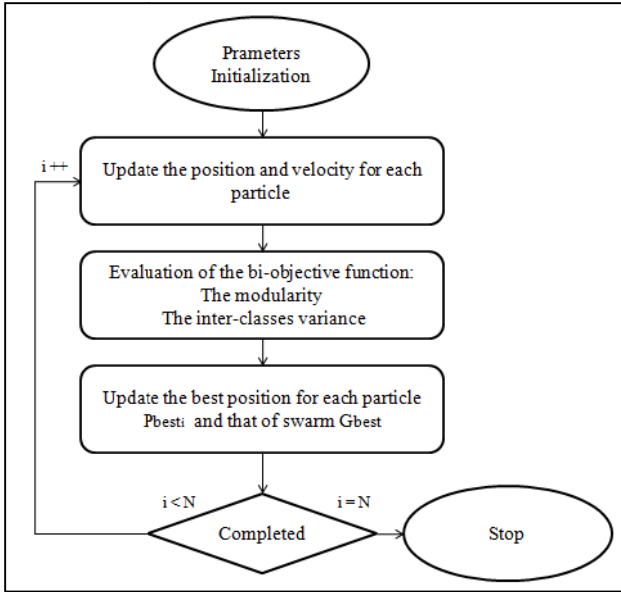


Fig. 1. The proposed community detection algorithm

4 Experimentation: Research Laboratory Case Study

To experiment our community detection solution, we chose part of a computer sciences research laboratory as an organization that contains 24 members. Then we implement the proposed contribution in a mixed decision support process that we introduced in [29]. The idea is when a participant within the organization has a problem, we use community detection to identify to this problem holder appropriate participants within the same organization that can be recommended to help him in the decision making process.

For the selected organization, modeled by the graph of fig.2, nodes represent the different members and edges represent the co-authoring relationships (i.e. common publications) among these researchers. To calculate the distance or similarity between nodes, we propose to model each member using an ontological profile introduced in [35]. This profile contains the following concepts: personal characteristics (Name, Age, Sex, and Nationality), competences, Profession and degrees for each member.

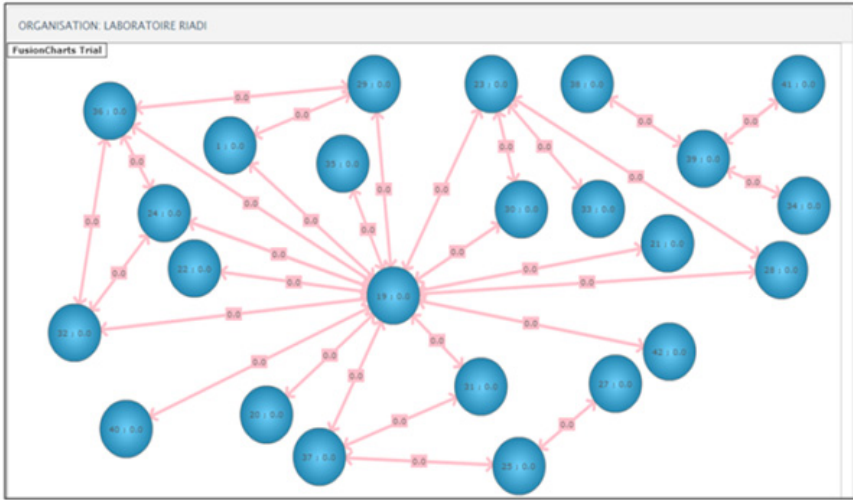


Fig. 2. The graph representing the selected organization

For each edge between i and j , the weight A_{ij} represents the similarity between the profiles of participants represented respectively by node i and node j . Then, we consider one member as problem holder and we will use the proposed algorithm to determine the community that can help this member to solve his/her problem. For each node j , the weight represents the similarity between this node profile and the profile of the problem holder.

In order to implement the profile ontology we use Protégé [36] to generate the owl file containing its concepts and their instances that represent the 24 members of the organization.

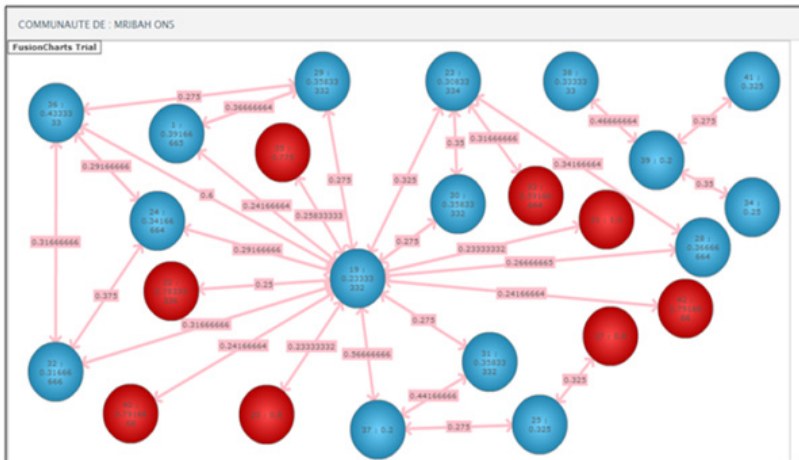


Fig. 3. The output community members (red nodes)

Then, we use JCOLIBRI framework [37], that offers different and predefined functions, allow the calculus of similarities between ontology instances so it facilitates the calculus of nodes and edges weights.

After calculating the similarities, we use PSO to find the threshold that maximizes the bi-objective function. For the selected example, the best particle illustrated by the threshold 0.33 give us the best fitness illustrated by the bi-objective quality 0.55. Using the Newman algorithm, the modularity is 0.3.

Finally, we display the output community members (represented by red nodes as it is seen in fig.3) recommended for help in the collaborative problem solving.

5 Conclusion

The aim of this paper is to evaluate community detection by taking into consideration two problematics. The first is to propose an objective function that considers the structure of the network, edges weights and nodes weights. The second is to maximize this function using an optimization technique as fast as possible which is particle swarm optimization.

References

1. Newman, M.E.J., Girvan, M.: Finding and Evaluating Community Structure in Networks. *J. Physical Review E* 69(I), 2 (2004)
2. Steinhaeuser, K., Chawla, N.V.: Community Detection in a Large Real-World Social Network. In: *International Conference on Social Computing, Behavioral Modeling and Prediction*, pp. 168–175 (2008)
3. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(8), 888–905 (2000)
4. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: Scan: A Structural Clustering Algorithm for Networks. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 824–833 (2007)
5. Tian, Y., Hankins, R.A., Patel, J.M.: Efficient Aggregation for Graph Summarization. In: *International Conference Management of Data (SIGMOD 2008)*, pp. 567–580 (2008)
6. Li, D.H., Liu, J.G., Liang, J.Z., Pana, Y.: Detecting Community Structure in Complex Networks Via Node Similarity. *Physica A: Statistical Mechanics and its Applications* 389(14), 2849–2857 (2010)
7. Zhou, Y., Cheng, H., Yu, J.X.: Graph Clustering Based on Structural/Attribute Similarities. In: *International Conference VLDB 2009*, pp. 718–729 (2009)
8. Dang, T.A., Viennet, E.: Community Detection Based on Structural and Attribute Similarities. In: *International Conference on Digital Society (ICDS 2012)*, pp. 7–14 (2012)
9. Wakita, K., Tsurumi, T.: Finding Community Structure in Mega-Scale Social Networks. In: *16th International Conference on World Wide Web (WWW 2007)*, pp. 1275–1276 (2007)
10. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast Unfolding of Communities in Large Networks. *J. of Statistical Mechanics: Theory and Experiment* 10, 10008–10020 (2008)

11. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature* 435(7043), 814–818 (2005)
12. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities Reveal Multiscale Complexity in Networks. *Natur.* 466(7307), 761–764 (2010)
13. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical Comparison of Algorithms for Network Community Detection. In: *International Conference on World Wide Web*, pp. 631–640 (2010)
14. Pizzuti, C.: Community Detection in Social Networks with Genetic Algorithms. In: *Genetic and Evolutionary Computation Conference* (2008)
15. Mazur, P., Zmarzłowski, K., Orłowski, A.J.: Genetic Algorithms Approach to Community Detection. In: *4th Polish Symposium on Econo- and Sociophysics* (2009)
16. Tasgin, M.: Community Detection Model Using Genetic Algorithm in Complex Networks and Its Application in Real-Life Networks, MS Thesis, Graduate Program in Computer Engineering, Bogazici University (2005)
17. He, D., Liu, J., Liu, D., Jin, D., Jia, Z.: Ant Colony Optimization for Community Detection in Large-Scale Complex Networks. In: *Seventh International Conference on Natural Computation (ICNC)*, pp. 1151–1155 (2011)
18. Sadi, S., Etaner-Uyar, S., Gündüz-Öğüdücü, S.: Community Detection Using Ant Colony Optimization Techniques. In: *15th International Conference on Soft Computing* (2009)
19. Liu, Y., Wang, Q., Wang, Q., Yao, Q., Liu, Y.: Email Community Detection Using Artificial Ant Colony Clustering. In: Chang, K.C.-C., Wang, W., Chen, L., Ellis, C.A., Hsu, C.-H., Tsoi, A.C., Wang, H. (eds.) *APWeb/WAIM 2007 Ws. LNCS*, vol. 4537, pp. 287–298. Springer, Heidelberg (2007)
20. Chen, B., Chen, L., Chen, Y.: Detecting Community Structure in Networks Based on Ant Colony Optimization. In: *International Conference on Information & Knowledge Engineering*, pp. 247–253
21. Guimera, R., Amaral, L.A.: Functional cartography of complex metabolic networks. *Nature* 433, 895–900 (2005)
22. El Dor, A.: Perfectionnement des algorithmes d’Optimisation par Essaim Particulaire. Applications en segmentation d’images et en électronique. PhD Thesis, Paris-Est University (2012)
23. Clerc, M., Siarry, P.: Une nouvelle métaheuristique pour l’optimisation difficile: la méthode des essaims particuliers. *J. l’enseignement des Sciences et Technologies de l’information et des Systèmes* 3(7) (2004)
24. Porter, M.A., Onnela, J.P., Mucha, P.J.: Communities in Networks. *Notices of the AMS* 56(9), 1082–1097 (2009)
25. Pons, P.: Détection de communautés dans les grands graphes de terrain. PhD Thesis, Paris 7 University (2007)
26. Amaral, L.A.N., Scala, A., Barthélémy, M., Stanley, H.E.: Classes of small-world networks. *Natl. Acad. Sci.* 97(21) (2000)
27. Lebart, L., Maurineau, A., Piron, M.: *Traitement des données statistiques*. Dunod, Paris (1982)
28. Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function. *J. of Amer. Statist. Assoc.* 58, 236–244 (1963)
29. Ben Yahia, N., Bellamine, N., Ben Ghezala, H.: Using Community Detection to Support Decision Making Process. In: *International Conference on Information Technology and e-Services*, pp. 566–570 (2012)
30. Eberhart, R.C., Kennedy, J.: New Optimizer Using Particle Swarm Theory. In: *6th International Symposium on Micro Machine and Human Science*, pp. 39–43 (1995)

31. Parsopoulos, K.E., Vrahatis, M.N.: Particle Swarm Optimization and Intelligence: Advances and Applications. Information Science Reference, IGI Global (2010)
32. Yisu, J., Knowles, J., Hongmei, L., Yizeng, L., Kell, D.B.: The Landscape Adaptive Particle Swarm Optimizer. *Applied Soft Computing* 8, 295–304 (2008)
33. Clerc, M.: The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization. In: *Congress of Evolutionary Computation*, pp. 1951–1957 (1999)
34. Carlisle, G., Dozier, G.: An Off-The-Shelf PSO. In: *Particle Swarm Optimization Workshop*, pp. 1–6 (2001)
35. Ben Yahia, N., Bellamine, N., Ben Ghezala, H.: Vers une architecture multicouche d'ontologies dédiée à la résolution mixte de problèmes. In: *Extraction et Gestion des Connaissances*, pp. 263–268 (2013)
36. <http://protege.stanford.edu>
37. Recio-García, J.A.: jCOLIBRI CBR Framework, <http://www.iccbr.org/iccbr10/jCOLIBRI-Overview.pdf>

New Algorithm for Solving Nonlinear Equations Roots

Delong Guo¹, Yongquan Zhou², and Xiaobin Luo¹

¹ Department of Maths Qiannan Normal College for Nationalities Duyun China
guodelong19760319@126.com

² College of Information Science and engineering,
Guangxi University for Nationalities Nanning China
Yongquanzhou@126.com

Abstract. Traditional optimization algorithm is widely used solving nonlinear equations numerical solution problem ,it is not only slow convergence speed but also easy to fall into local optimal solution and solution low precision. Adaptive membrane computing optimization algorithm is important achievement performance improvement, Firstly, the high-dimensional space split,each subspace is a basic membrane, evolutionary strategy algorithm based on basic membrane area is used to improve the local search ability and convergence speed. Basic membrane area will be local optimum timing is transmitted to the surface membrane. Particle swarm optimization (PSO) has global search ability is used surface membrane area. through simulation the paper can comparatively analyze the performance of different algorithms.

Keywords: membrane computing, evolutionary strategy, global optimization, regional.

1 Introduction

Many of the mathematical physics problems often are attributed to the solutions of nonlinear equation. The nonlinear equation of root is one of the important practical significance of the problem, but also many mathematicians had being committed to solve the problem for many years. For a long time,many methods are found and used to solve nonlinear equation of roots. Newton's method and string cut method and parabolic method [1, 2] are traditional methods .Then, many newton method derived from algorithms were proposed , such as newton-down method, etc. These traditional methods are dependent on the initial value .for the nonlinear equation of problem, it is necessary to explore a kind of high efficient and reliable method. The European academy of sciences, Romania scientists Gheorghe Paun is propoesd Membrane Computing (Membrane Computing, MC) in 1998 [3], also called P system (P Systems). It is a whole new area of research in computer science, it was proposed soon the research focus of many scholars, many membrane computing models were applied to the approximate optimization, computer graphics, cryptography, parallel computing, and other fields. In the membrane structure, the rules of each membrane region is run independently membrane exchanges between only occur between the adjacent two

regions, the segmentation algorithm is easily in parallel, distributed or grid computing system implemented. For other optimization algorithms (such as: differential evolution algorithm, genetic algorithm, etc.) in low dimensional unimodal optimization problem is strong optimization ability, but for high dimension large space optimization problems, along with the increase of the dimension, problem solving is difficulty greatly increased. For example optimization problem which is multimode problem has more than one local optimization points, it is more complex and difficult to obtain the global optimal solution. In view of traditional optimization algorithm is combined with membrane computing, through the membrane and membrane communication between gradually to search the global optimal solution. In this paper, according to membrane algorithm, based on the membrane of high dimension global optimization algorithm, membrane structure is divided into two layers: basic membrane (Elementary Membrane) and surface membrane (Skin Membrane), the algorithm first will be a high dimensional space for segmentation which is a basic membrane. the evolution strategy is used basic membrane area in the local search strategy which is to improvement the algorithm of local convergence speed and search ability. Basic membrane optimum solution is passed to the surface membrane, PSO is used surface membrane region .

2 Membrane Computing the Step

- (1) Update: sub algorithm is used to update solution(solutions) in each region at the same time (parallel) .
- (2) transmission exchange: respectively each area optimization of the best solution (solutions) and worst solution (solution) is sent to the adjacent upper (father layer) and low (sublayer).
- (3) repeat (1), (2) the steps until termination condition is satisfied.

3 The Description of the Problem

A nonlinear equation $f(x) = 0$ (1) all the real roots are included in the interval (c, d) . definition a real root $x^* \in [a, b]$ is a fixed solution of only equation (1) .

$[a, b] \subset [c, d]$ the optimization theory easy proof equation (1) in $[a, b]$ the root of the equivalent to function $F(x) = (f(x))^2$ (2)

In $[a, b]$ the minimum point is one of equation (1) exact solution, if and only if $F(x)$ minimum value is 0.

The minimum point is one of equation (1) approximate solution, if and only if $F(x)$ the minimum value is satisfied with precision .

4 Solving Nonlinear Equation of Root Based on the Membrane Computing Optimization Algorithm

- (1) space division: Using evolutionary strategy algorithm of feasible space $[L, U]$ is divided into m mutually disjoint $[L_s(i), u_s(i)], i = 1, 2, \dots, m$. Each subspace $[L_s(i), u_s(i)]$ as membrane system of a basic membrane area M .
- (2) population initialization: In each basic membrane area M_i randomly generated initialization sub-population.
- (3) membrane local search: In each basic membrane area M_i evolutionary strategy algorithm is used local search, if the local optimal solution to get updates, which is sent to a surface membrane area.
- (4) global search: the improved adaptive learning PSO algorithm is used in the surface area of membrane.
- (5) termination condition judgment: If achievement the required number of iterations or be satisfied with the results, ends and outputs the result otherwise go to (3).

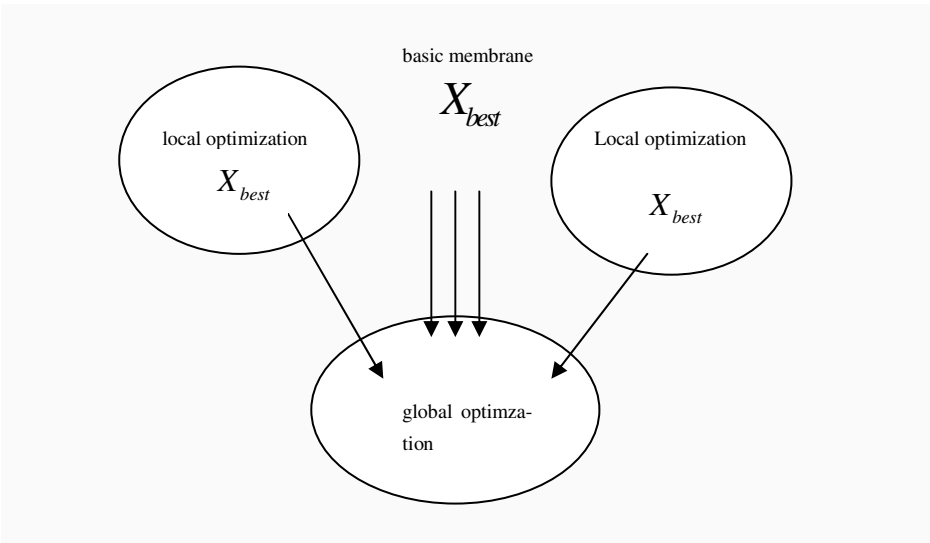


Fig. 1. This paper algorithm membrane structure

5 Numerical Simulation

Example 1 to test the performance of the algorithm in this paper, the selection of reference [9-11] 5 a linear equations

1. $f(x) = x^3 - 2x - 5 = 0, x \in [-4, 4]$
2. $f(x) = x^3 - 3x^2 - 6x + 8 = 0, x \in [0, 2]$

- 3. $f(x) = xe^x - 1 = 0, x \in [0, 4]$
- 4. $f(x) = x^4 - 12x^3 + 47x^2 - 60x = 0, x \in [-10, 10]$
- 5. $f(x) = (x - 1)^2 \sin^2(x) + (x - 1)^3 \cos^3(x) + 5(x - 1) = 0, x \in [-5, 5]$

Numerical simulation in the Matlab2010a finish, membrane system is composed of five basic membrane and 1 surface membrane composition. Each membrane the area of the initial population set maximum 40, surface membrane area PSO algorithm parameters Settings for example 1 to test the performance of the algorithm in this paper, the selection of reference [10-12] 5 a linear equations $L(0) = 0.1, b = 1, v_{id} = 0.1, i = 1, 2, \dots, popsize; d = 1, 2, \dots, N, N$ is function dimension.. search algorithm parameters respectively the number of initial population $\mu = 50, \lambda = 350 \sigma(0) = 3.0$ r' global coefficient, often take 1; r - local coefficient, often take 1

Table 1. This paper algorithm and the references [9] for example 1 1-3 the comparison of the problem

algorithm.	success rate	x mean value	$f(x)$ mean value	precision
1 references[1]	100%	2.094554	0.000000	10^{-6}
this algorithm	100%	2.094551231	0.0000000	10^{-12}
2 references[9]	100%	0.999999	0.000000	10^{-6}
this algorithm	100%	1.000000000	0.0000000	10^{-12}
3 references[9]	100%	0.567156	0.000000	10^{-6}
this algorithm	100%	0.5671426890	0.0000000	10^{-12}

Be seen from the comparison results in Table 1, the membrane algorithm proposed in this paper than references[10] in the algorithm has a higher accuracy. references[10] algorithm is a single particle swarm algorithm, this algorithm is a reasonable combination of a wide range of particle swarm algorithm to search for global solutions and evolutionary strategy algorithm local meticulous search solution characteristics.

Table 2. Example 1 fourth problems in case of numerical simulation results

precision solution	x mean value	$f(x)$ mean value	precision	success rate
0	0.000000000000	0.000000000000	10^{-13}	100%
3	3.000000000000	0.000000000000	10^{-13}	100%
4	4.000000000000	0.000000000000	10^{-13}	100%
5	5.000000000000	0.000000000000	10^{-13}	100%

For example 4, reference[11] interval [4,5,6] the range found a precision for roots in 5.0026, and this algorithm in the expanded to [-10,10] the case of successful found example 4 total solution, and the precision of reach 10^{-13}

Table 3. Example 1 fifth problem of numerical simulation results

algorithm.	success rate	X mean value	$f(x)$ mean value	precision
references[11]	100%	-4.969225181	0.000000000	10^{-10}
	100%	-2.294379447	0.000000000	10^{-10}
	100%	3.275770293	0.000000000	10^{-10}
this algorithm	100%	-4.96922517156	0.00000000000	10^{-12}
	100%	-2.29437943531	0.00000000000	10^{-12}
	100%	1.00000000000	0.00000000000	10^{-12}
	100%	3.27576018442	0.00000000000	10^{-12}
	100%	3.43358867531	0.00000000000	10^{-12}

From the results in Table 3 show, references [12] found only 5 cases of 3 solutions, and this algorithm has successfully found it all 5 solutions, and accuracy than references [12] high, shows that this algorithm has stronger versatility
 For example 2 Select 3 multivariate nonlinear equations

$$\begin{aligned}
 x_1^3 + x_1^2 x_2^2 x_3 + x_1 x_2 x_3^3 + 2 x_2 &= 3 \\
 x_1^5 x_2^4 + x_1 x_2 x_3^2 + x_2^2 x_3 + 3 x_1 &= 4 \\
 x_3^5 + x_1^2 x_2^2 + x_1 x_3 + 3 x_1 &= 4
 \end{aligned}
 \tag{3}$$

Each membrane the area of the initial population set maximum 40, surface membrane area
 PSO algorithm parameters Settings
 $L(0) = 0.1, b = 1, v_{id} = 0.1, i = 1, 2, \dots, popsiz$; $d = 1, 2, \dots, N, N$ is function dimension..
 search algorithm parameters respectively the number of initial population $\mu = 30, \lambda = 210 \sigma(0) = 3.0$ r' global coefficient, often take 1; r - local coefficient, often take 1

Table 4. The algorithm and other algorithms by solving this equation group

algorithm	convergence Solutions	Initial value	iteration
ABS algorithm[12]	(1.000001, 0.999999, 0.000008)	(0.85, 0.85, 0.15)	8
PSO algorithm[13]	(0.999999627, 1.000000650, 0.000000134)	randomly given	127
This algorithm	(0.999999736, 1.000000648, 0.000000133)	randomly given	115

From Table 4 we can see this algorithm is better than the traditional ABS algorithm, this algorithm is not affected by the display of choosing the initial value of influence, even in are randomly choosing the initial value condition calculation speed is faster than PSO algorithm, but also has high precision, effectively avoids the difficult problem of initial point selection.

6 Conclusions

This paper presents a new algorithm based on membrane computing optimization which makes full use of the characteristics of membrane structure of the particle swarm optimization (PSO) global convergence advantages and evolutionary strategy for local search to combine the advantages of application in solving nonlinear equations, numerical solution is proposed. From the numerical calculation results can be seen in this design the algorithm not only has good stability and high computing speed and accuracy, it is a feasible and effective optimization algorithm and for the algorithm to other applications and provide a new way.

References

1. Brown, L.: A Survey of Image Registration Techniques. *ACM Computer Surveys* 24(4), 325–376 (1992)
2. Li, Q.Y., Wang, C.N., Yi, D.Y.: *Numerical analysis*, pp. 2212–2232. Tsing University Publishing Press, Wuhan (2008)
3. Paun, G.: Computing with membrane. *International Journal of Foundations of Computer Science* 11(1), 167–182 (2000)
4. Busi, N.: Using well-structured transition systems to decide divergence for catalytic P systems. *Theoretical Computer Science* 372(2/3), 125–135 (2007)
5. Zhang, G.X., Pan, L.: Nature computation new branch-membrane computation. *Journal of Computer* 33(2), 208–214 (2010)
6. Nishida, T.: An application of P system: a new algorithm for NP-complete optimization problems. In: *Proceedings of 8th World Multi-Conference on Systems, Cybernetics and Information*, pp. 109–112 (2004)
7. Nishida, T.: An approximate algorithm for NP-complete optimization problems exploiting P systems. In: *Proceedings of Brainstorming Workshop on Uncertainty in Membrane Computing*, Palma de Majorca, pp. 185–192 (2004)
8. Yun, Q.X.: *Evolutionary algorithm*. Metallurgical Industry Press, BeiJing (2000)
9. Zhang, J.K.: Solving nonlinear equation and equation of particle swarm optimization (ps) algorithm. *Computer Engineering and Application* 42(7), 56–58 (2006)
10. Zang, M.L., Yang, S.J.: Using genetic algorithm avoids solving nonlinear equation. *Jining College Journal* 19(6), 1–2 (1998)
11. Gao, F., Yan, H.Q.: A kind of solving equation root of improved particle swarm optimization algorithm. *Journal of Wuhan University (Physical Edition)* 52(3), 296–300 (2006)
12. Ge, R.D., Xia, Z.Q.: A modified solution to a class of singular nonlinear system of equations of ABS algorithm. *Journal of Dalian University of Technology* 43(6), 704–710 (2003)
13. Sun, M.J., Chen, Y.X., Hu, Q.: Solving singular nonlinear system of equations of particle swarm optimization algorithm. *Journal of Heilongjiang Institute of Science and Technology* 16(6), 369–373 (2006)

Relationship between Blackwell's Informativeness and Supermodular Precision

Yuanyuan Li*

Bielefeld University and University Paris 1-Pantheon Sorbonne
leedoubleyuan@gmail.com, yuan_yuan.li@uni-bielefeld.de

Abstract. In a decision problem with uncertainty the decisions are supposed to be made after obtaining information from some signals, thus the expectation conditional on signals is likely to be crucial. Therefore, the relationship between the classical Blackwell's informativeness criterion and the dispersion of conditional expectations, one of which is the so-called supermodular precision criterion, is examined here. The study shows that Blackwell's informativeness does not necessarily imply or be implied by the supermodular precision in general cases even with the monotone likelihood ratio property, although binary cases could enlighten us somehow.

Keywords: Blackwell's Informativeness, Dispersive orders, Posterior distributions, Supermodular Precision.

1 Introduction

Uncertainty exists nearly everywhere and the true state is often unknown to agents, yet in most cases agents can overcome the uncertainty by obtaining related information that is conveyed in some signals. Their decisions will then be made based on the revised knowledge, i.e. the posterior belief of the state.

But how would information influence the decision? Various informativeness criteria have been proposed during the last decades, while until very recently a very intuitive idea is introduced by Ganuza and Penalva (2010), following the intuition that more information should lead to a more disperse conditional expectations, and a new type of criteria, which they refer to as precision criteria, are defined in their paper.

Although the formalization of precision criteria provides an easier way to interpret the informativeness of signals, the ordering by the precision criteria is, unconventionally, based on conditional expectations rather than directly on the underlining information systems. So here follows the questions: could it be possible to link the precision criteria with the traditional informativeness criteria? Is there some implications among these criteria? Therefore, to bridge the gap between the informativeness and the dispersion criteria is one of the main purposes of this study.

* Under the program of European Doctorate in Economics Erasmus Mundus (EDEEM).

2 Informativeness and Dispersive Orders

2.1 Information Systems

In the setting of statistical decision theory, uncertainty can be characterized by a random state of nature. As stated before, a decision maker takes actions before knowing the true state but after observing a signal which is correlated to the true state. We further assume that the agent has a prior probability distribution of the state of nature, then he can infer additional information from the signal and revise his belief about the true state via Bayes' Rule. In this context, an information system is used to formalize how the signals are generated by the true state.

Definition 1. The triplet (Ω, Y, F) is defined as an information system, where Ω is the set of states, Y is the set of signals, and F is a stochastic transformation from Ω to Y , represented by the conditional density function $f(y|\omega)$, for any $y \in Y$ and $\omega \in \Omega$.

In the finite discrete case, suppose that there are N possible states, i.e. $\Omega = \{\omega_1, \dots, \omega_N\}$, and M different signal values, i.e. $Y = \{y_1, \dots, y_M\}$. Then, the transformation F can be represented by a matrix with the entry f_{ij} showing the probability of generating a signal y_j under the state ω_i , where $i \in \{1, 2, \dots, N\}$ and $j \in \{1, 2, \dots, M\}$.

Then, for a given prior $\pi(\cdot)$ defined on Ω , agents can update their beliefs; the posterior belief is given by:

$$v(\omega|y) = \frac{f(y|\omega)\pi(\omega)}{\mu(y)}, \quad \forall \omega \in \Omega, \forall y \in Y, \quad (1)$$

where $\mu(y) = \int_{\Omega} f(y|\omega')\pi(\omega')d\omega'$. Decisions then will be made according to the posterior beliefs.

The availability of additional information allows agents to better react to the risky environment, which leads naturally to the consideration of the value of information, or of a certain information system. Consider an expected utility maximizer with a utility function $u(a, \omega)$, where $a \in A$ is the action the agent takes. After receiving a signal y , the agent chooses the optimal action $a^*(y)$ which solves the maximization problem $\max_{a \in A} \int_{\omega \in \Omega} u(a, \omega)v(\omega|y)d\omega$. And the value of the information is

$$V := \int_{y \in Y} \mu(y) \int_{\omega \in \Omega} u(a^*(y), \omega)v(\omega|y)d\omega dy. \quad (2)$$

That is, the value of information is the agent's ex ante expected utility from an optimal chosen decision rule. Then, it is feasible to compare two different information systems for a given decision problem.

2.2 Blackwell's Informativeness

When there are more than one way of obtaining information, we would consider the problem of how to choose a better one. For this purpose various criteria have been proposed, among which a classical criterion was developed by Blackwell (1953). It follows the intuition that a less informative system can be duplicated from a more informative one by adding some random transmission error. Formally,

Definition 2. Let (Ω, Y^F, F) and (Ω, Y^G, G) be two information systems. Then (Ω, Y^F, F) is said to be more informative than (Ω, Y^G, G) , denoted by $F \succ_i G$, if there exists a stochastic transformation Γ from Y^F to Y^G represented by a stochastic density kernel γ such that for all $\omega \in \Omega$ and $y^G \in Y^G$, it holds true that

$$g(y^G | \omega) = \int_{y^F \in Y^F} \gamma(y^G, y^F) f(y^F | \omega) dy^F. \quad (3)$$

For the finite case, information systems can be represented with matrices, thus Definition 2 can be modified by replacing Equation (3) with the following condition:

$$G = F \Gamma \quad (4)$$

where Γ is a stochastic matrix.

As a matter of fact, a more informative system, as defined above, is indeed preferred by every expected utility maximizer, i.e. $V^F \geq V^G$ for all utility functions u , as stated in the well-known Blackwell's Theorem.

2.3 Orders of Dispersion

Intuitively, signals containing more useful information would have a stronger impact on posterior distributions, which may in turn lead to more dispersive conditional expectations. Thus, Ganuza and Penalva (2010) introduce a new kind of criteria for evaluating different information systems by applying different stochastic orders to conditional expectations of the states, which they call the precision criteria. Two criteria - the supermodular precision and the integral precision - based on two different orders are focused in their study. For instance, the supermodular precision is defined on the basis of the dispersive order, which is also one of most important dispersion criterion we will focus on later.

Definition 3. X and Y are two real-valued random variables with distributions F_X and G_Y , respectively. Then, X is said to be greater than Y in the dispersive order (denoted by $X \succeq_{disp} Y$), if for all $0 < p < q < 1$,

$$F_X^{-1}(q) - F_X^{-1}(p) \geq G_Y^{-1}(q) - G_Y^{-1}(p) \quad (5)$$

where F_X^{-1} and G_Y^{-1} are quantile functions of X and Y , respectively.

Rearranging equation (5), we can directly get the following characterization of the dispersive order.

Lemma 1. $X \geq_{disp} Y$ if, and only if, $F_X^{-1}(p) - G_Y^{-1}(p)$ is increasing in $p \in (0,1)$.

Now we could introduce Ganuza and Penalva’s criterion based on the dispersive order to describe the dispersion property of the conditional expectations. Let $\tilde{\omega}$ be the **random** variable of the state and \tilde{y}^F and \tilde{y}^G be the random variables of the signals generated from information system (Ω, Y^F, F) and (Ω, Y^G, G) , respectively¹. Then, we can define the so-called supermodular precision criterion as follows.

Definition 4. An information system (Ω, Y^F, F) is said to be more supermodular precise than another system (Ω, Y^G, G) , denoted by $F \succ_{sm} G$, if $E[\tilde{\omega} | \tilde{y}^F] \geq_{disp} E[\tilde{\omega} | \tilde{y}^G]$.

Similar as the supermodular precision criterion, the integral precision is also defined on conditional expectations but with another stochastic order - the convex order. Therefore, some other orders may also be considered to define the precision of information systems, such as mean-preserving spread, single-crossing dispersion, rotation, etc.

3 Relationship between Informativeness and Supermodular Precision

3.1 Binary Examples

Example 1. There are two types of projects, namely, low- and high-revenue projects, with payoffs of q_L and q_H , respectively, where $q_L < q_H$. Assume that there are a fraction π of the low-revenue projects and $1-\pi$ fraction of high-revenue ones, with $0 < \pi < 1$. The investment decision is made after observing a payoff-related signal $\tilde{y} \in \{y_L, y_H\}$. Let $p_t := Pr(y_t | q_s)$ where $t, s \in \{L, H\}$ and $t \neq s$. Then, p_t can be interpreted as the “error probability”. Furthermore, we assume that $p_L = p_H = p \in [0, \frac{1}{2})$. Therefore, the information system is characterized by $(F, \{q_L, q_H\}, \{y_L, y_H\})$, where

$$F = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

Lemma 2. For the binary case stated as above, the larger the error probability p is, the smaller the difference between the conditional expected payoffs are. That is, $E[\tilde{q} | y_H] - E[\tilde{q} | y_L]$ is decreasing in p .

Proof. Note that the probability of receiving a low signal is given by $P(y_L) = P(q_L)P(y_L | q_L) + P(q_H)P(y_L | q_H) = \pi(1-p) + (1-\pi)p$, then we have

¹ Without any other specification, we will use in the paper the tilde sign to represent the corresponding random variables.

$$P(q_L | y_L) = \frac{P(q_L)P(y_L | q_L)}{P(y_L)} = \frac{\pi(1-p)}{\pi(1-p) + (1-\pi)p}$$

and

$$P(q_H | y_L) = \frac{P(q_H)P(y_L | q_H)}{P(y_L)} = \frac{(1-\pi)p}{\pi(1-p) + (1-\pi)p}$$

Thus, we have

$$\begin{aligned} E[\tilde{q} | y_L] &= P(q_L | y_L)q_L + P(q_H | y_L)q_H \\ &= \frac{\pi(1-p)}{\pi(1-p) + (1-\pi)p} q_L + \frac{(1-\pi)p}{\pi(1-p) + (1-\pi)p} q_H. \end{aligned}$$

Similarly, the expected utility of the payoff conditional on the high signal is given by

$$E[\tilde{q} | y_H] = \frac{\pi p}{\pi p + (1-\pi)(1-p)} q_L + \frac{(1-\pi)(1-p)}{\pi p + (1-\pi)(1-p)} q_H$$

Denote $\Delta E := E[\tilde{q} | y_H] - E[\tilde{q} | y_L]$, and taking derivatives of this difference with respect to the error probability p yields that

$$\frac{d\Delta E}{dp} = \frac{\pi(1-\pi)q_L + \pi(\pi-1)q_H}{[\pi p + (1-\pi)(1-p)]^2} - \frac{\pi(\pi-1)q_L + \pi(1-\pi)q_H}{[\pi(1-p) + (1-\pi)p]^2} < 0$$

where the inequality follows the fact that $q_H > q_L$. QED.

In another word, Lemma 3 shows that for any $0 < p < p' < 1/2$,

$$E^F[\tilde{q} | y_H] - E^F[\tilde{q} | y_L] > E^G[\tilde{q} | y_H] - E^G[\tilde{q} | y_L], \quad (7)$$

where E^F and E^G represent the expectation calculated under the information systems with parameters p and p' , respectively. Equation (7) exhibits a property of the dispersive order, although the necessary condition of $E^F[\tilde{q} | \tilde{y}] \geq_{disp} E^{F'}[\tilde{q} | \tilde{y}]$ requires that the probability of $E^F[\tilde{q} | \tilde{y} = y_L]$ and $E^{F'}[\tilde{q} | \tilde{y} = y_L]$ are equal, which is also equivalent to $Pr^F(\tilde{y} = y_L) = Pr^{F'}(\tilde{y} = y_L)$. Therefore, to draw the conclusion of the dispersion of the conditional expectations we need in addition that

$$\pi(1-p) + (1-\pi)p = \pi(1-p') + (1-\pi)p'.$$

In fact, the equal probability property together with Equation (7) is also sufficient to show the dispersive ordering. Hence, we know that for some prior distribution $p < p'$ implies that $E^F[\tilde{q} | \tilde{y}] \geq_{disp} E^G[\tilde{q} | \tilde{y}]^2$. That is, the information system F is more supermodular precise than the information system G . Nevertheless, in this example the only possible prior distribution is $\pi=1/2$.

² Moreover, if the investment decision is made on the basis of the expected utilities of the possible payoffs, we can also have the same conclusion that a higher error probability results in a smaller dispersion of conditional expected utilities as long as the utility function is increasing.

Now we consider some general binary information systems.

Example 2. There are two information systems which are characterized by $(\{q_L, q_H\}, \{y_L, y_H\}, F)$ and $(\{q_L, q_H\}, \{y_L, y_H\}, G)$, respectively, where

$$F = \begin{pmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} 1-q_1 & q_1 \\ q_2 & 1-q_2 \end{pmatrix}$$

with $p_1, p_2, q_1, q_2 \in (0,1)$, $p_1 + p_2 < 1$ and $q_1 + q_2 < 1$.

Proposition 1. *F and G are two information systems as stated above, with the prior distribution. Then, $F \succ_i G$ implies $F \succ_{sm} G$ for some prior $0 < \pi < 1$. That is, in the binary cases, the more informative the information system is, the more dispersive the conditional expectation is.*

Proof. The conditional expectations in the more informative information system are given by

$$E^F[\tilde{q} | y_L] = \frac{\pi(1-p_1)q_L + (1-\pi)p_2q_H}{\pi(1-p_1) + (1-\pi)p_2}, \quad E^F[\tilde{q} | y_H] = \frac{\pi p_1q_L + (1-\pi)(1-p_2)q_H}{\pi p_1 + (1-\pi)(1-p_2)}$$

For the less informative one, we have

$$E^G[\tilde{q} | y_L] = \frac{\pi(1-q_1)q_L + (1-\pi)q_2q_H}{\pi(1-q_1) + (1-\pi)q_2}, \quad E^G[\tilde{q} | y_H] = \frac{\pi q_1q_L + (1-\pi)(1-q_2)q_H}{\pi q_1 + (1-\pi)(1-q_2)}$$

Thus, we have the difference of the expected payoffs conditional on signal y_L in the two different information systems

$$\begin{aligned} E^F[\tilde{q} | y_L] - E^G[\tilde{q} | y_L] &= \frac{\pi(1-p_1)q_L + (1-\pi)p_2q_H}{\pi(1-p_1) + (1-\pi)p_2} - \frac{\pi(1-q_1)q_L + (1-\pi)q_2q_H}{\pi(1-q_1) + (1-\pi)q_2} \\ &= \frac{\pi(1-\pi)[q_H - q_L][p_2(1-q_1) - (1-p_1)q_2]}{[\pi(1-p_1) + (1-\pi)p_2][\pi(1-q_1) + (1-\pi)q_2]} \leq 0 \end{aligned}$$

since $p_2(1-q_1) \leq (1-p_1)q_2$, which derives from the assumption that $F \succ_i G$. Thus, we have $E^F[\tilde{q} | y_L] \leq E^G[\tilde{q} | y_L]$. Similarly, we know $E^F[\tilde{q} | y_H] \geq E^G[\tilde{q} | y_H]$. Therefore, we can conclude that $E^F[\tilde{q} | y_H] - E^F[\tilde{q} | y_L] \leq E^G[\tilde{q} | y_H] - E^G[\tilde{q} | y_L]$.

Then, for the prior $0 < \pi < 1$ such that $Pr^F(\tilde{y} = y_L) = Pr^G(\tilde{y} = y_L)$, i.e. $\pi(1-p_1) + (1-\pi)p_2 = \pi(1-q_1) + (1-\pi)q_2$, we have that the conditional expectations in the more informative system are more dispersed compared to the less informative one. QED.

Although the informativeness criterion implies the supermodular precision criterion which is defined by the dispersion of the conditional expectations, the implication also depends on the prior, which ensures one of the necessary conditions of the dispersive order.

3.2 General Discrete Cases

Now we consider a more general case where $\#Q = n$ and $\#Y = m$, the result becomes unclear. Let $Q = \{q_1, \dots, q_n\}$, sorted in increasing order, be a finite set of possible payoffs, with the prior probability $\{\pi_1, \dots, \pi_n\}$; $Y = \{y_1, \dots, y_m\}$ is the set of signals, which is also sorted in increasing order. Two information systems are identified by the following matrices, F and G , respectively.

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & & \vdots \\ f_{n1} & \cdots & f_{nm} \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} g_{11} & \cdots & g_{1m} \\ \vdots & & \vdots \\ g_{n1} & \cdots & g_{nm} \end{pmatrix}.$$

Suppose that F is more informative than G in Blackwell's sense, denoted as $F \succ_i G$. That is, there exists a stochastic matrix $\Gamma = (\gamma_{ij})_{m \times n}$ such that

$$G = F \Gamma \tag{8}$$

Moreover, assume that both information systems have the monotone likelihood ratio property (MLRP)³.

Actually, the procedure of the proof above can be easily generalized to the finite discrete case. However, the implication between informativeness and the supermodular precision doesn't hold any more, even for those priors which yield the same jump of the quantile functions. An counter example is shown in the following.

Let $Q = \{q_1, q_2\}$ and $Y = \{y_1, y_2, y_3\}$, and the information systems are given by

$$F = \begin{pmatrix} 0.9195 & 0.0517 & 0.0288 \\ 0.1053 & 0.2399 & 0.6548 \end{pmatrix}, \text{ and } G = \begin{pmatrix} 0.7141 & 0.1974 & 0.0885 \\ 0.2489 & 0.1381 & 0.6130 \end{pmatrix}$$

We can check that both F and G satisfy MLRP and there exists a stochastic matrix Γ such that $G = F\Gamma$, i.e. $F \succ_i G$. However, even for the prior $(0.4113, 0.5887)$, which ensures the same probability of each signal realization, we have

$$E^F[\tilde{q} | y_3] - E^F[\tilde{q} | y_2] = 0.1010 < E^G[\tilde{q} | y_3] - E^G[\tilde{q} | y_2] = 0.4082.$$

In fact, the property at the extremes can be kept, while the monotonicity of $E^F[u(\tilde{q}) | y_s] - E^G[u(\tilde{q}) | y_s]$ with respect to y_s can be possibly violated in the intermediate values when there are more than two signal realization values. The MLRP guarantees the monotone increasing of the conditional expectations in both information system, yet the speed of the increasing can differ, which possibly makes the difference of the conditional expectations fluctuate.

³ The densities $\{f(y|\omega)\}_{y \in Y, \omega \in \Omega}$ have the monotone likelihood ratio property, if for every $y' > y$ and $\omega' > \omega$, $f(y'|\omega')f(y|\omega) - f(y'|\omega)f(y|\omega') \geq 0$.

Although the result cannot be duplicated for general information systems, it is true as long as the signal space contains only 2 elements, i.e. $m = 2$. Formally,

Proposition 2. *F and G are two information systems with finite states and 2 signals, and the prior distribution is (π_1, \dots, π_n) . Then, $F \succ_i G$ implies $F \succ_{sm} G$ for the prior $0 < \pi < 1$ such that $Pr^F(\tilde{y} = y_i) = Pr^G(\tilde{y} = y_i)$ for all $i \in \{1, 2\}$.*

That is, the conditional expectation is more dispersive under a more informative system even when the set of states contains more than two elements.

4 Conclusion

As we show above, Blackwell's sufficiency criterion implies the dispersion of conditional expectations in binary cases, though it does not necessarily have the relationship for general cases. However, the relationship with some other informativeness criteria could be built up under certain conditions.

Obviously, Blackwell's criterion is a partial order and does not have a large power to order information systems. For this reason, instead of considering all decision problem, one would sometimes focus only on some special problems, and therefore, be able to compare more information systems. Alternative criteria for comparing two different information systems have also been introduced. For example, Lehmann (1988) considers information systems with monotone likelihood ratio property, which is also known as the property of affiliation. Besides, Persico (2000) formalizes Lehmann's criterion, where the informativeness of an information structure is defined in the following way, where the term "accurate" is used⁴.

Therefore, further study could be made based on some other criteria, and our preliminary study shows that the informativeness captured by accuracy criterion does have some relationship with the dispersion of the expectations conditional on signals. This could provide theoretical basis for applying the dispersive orders directly to decision analysis.

References

1. Blackwell, D.: Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics* 24, 265–272 (1953)
2. Crémer, J.: A Simple Proof of Blackwell's "Comparison of Experiments" Theorem. *Journal of Economic Theory* 27, 439–443 (1982)
3. Eckwert, B., Zilcha, I.: Incomplete Risk Sharing Arrangements and the Value of Information. *Economic Theory* 21, 43–58 (2003)
4. Ganuza, J., Penalva, J.S.: Signal Ordering Based on dispersion and the Supply of Private Information in Auctions. *Econometrica* 78(3), 1007–1030 (2010)

⁴ More recently, Jewitt (2007) relates the Blackwell's criterion with Lehmann's and extends Lehmann's results in a general single crossing preference (SCP) case.

5. Jewitt, I.: Information Order in Decision and Agency Problems. Personal manuscript. Nuffield College, University of Oxford (2007)
6. Le Cam, L.: Sufficiency and Approximate Sufficiency. *Annals of Mathematical Statistics* 35, 1419–1455 (1964)
7. Lehmann, B.L.: Comparing Location Experiments. *Annals of Statistics* 16(2), 521–533 (1988)
8. Milgrom, P.R.: Good News and Bad News: Representation Theorems and Applications. *Bell Journal of Economics* 12, 380–391 (1981)
9. Persico, N.: Information Acquisition in Auctions. *Econometrica* 68(1), 135–148 (2000)

A Novel Genetic Programming Based Classifier Design Using a New Constructive Crossover Operator with a Local Search Technique

Arpit Bhardwaj and Aruna Tiwari

Indian Institute of Technology,
Computer Science Department, Indore, India
{phd12110102,artiwari}@iiti.ac.in

Abstract. A common problem in genetic programming search algorithms is the destructive nature of the crossover operator in which the offspring of good parents generally has worse performance than the parents. Designing constructive crossover operators and integrating some local search techniques into the breeding process have been suggested as solutions. In this paper, we proposed the integration of variants of local search techniques in the breeding process, done by allowing parents to produce many off springs and applying a selection procedure to choose high performing off springs. Our approach has removed the randomness of crossover operator. To demonstrate our approach, we designed a Multiclass classifier and tested it on various benchmark datasets. Our method has shown the tremendous improvement over the other state of the art methods.

Keywords: Genetic Programming, Crossover, Local Search Technique.

1 Introduction

Crossover (sexual recombination) is recognized as the primary genetic operator for improving program structures in tree-based (GP) [1]. It plays vital role in improving GP process. It plays a critical role in deriving optimal solutions as shown by the large number of studies related to crossover operators since the 1990s. The standard crossover operator [1] picks a random crossover point in each of two parent program trees, and swaps the two sub trees rooted at the chosen crossover points to generate two new programs. This has been seen as problematic: most crossover events in the standard crossover produce offspring with less than half of the fitness of their parents.

The crossover genetic operator has been considered “the center of the storm” in GP [2]. In the current crossover operator, two subprograms (crossover points) are randomly chosen from two parent programs, and two new programs are generated by simply swapping them. However, the totally random choice is clearly unable to guarantee the best choice. More importantly, this random choice often destroys the good “building blocks” (subprograms that are good for that task) in evolved programs, particularly in the later stage of evolution [2].

Selection pressure in the selection of parent programs is intended to improve the average problem solving quality of the population. It gives individuals of higher quality a higher probability of being used to create the next generation so that the search algorithm focuses on promising regions in the search space [3].

In the standard breeding process of the GP algorithm, crossover operators produce two offspring from a pair of parents programs. With the standard breeding process, exploring new states in the neighborhood search space of current states can be viewed as a set of random walks. If parents are randomly selected for mating, the GP algorithm will effectively act like a random (beam) search algorithm. Therefore selection pressure is applied to the parent selection process to reduce the stochastic element of the search and to provide individuals having good fitness with more chances to be chosen than others. Good genetic material in the chosen individuals is expected to be propagated along evolution in order to speed up population convergence.

With the standard breeding process, applying selection pressure on the parent selection has been recognized as more effective than selecting parents randomly. Individuals with good fitness are used with increased frequency to produce offspring. However, the number of possible offspring in the immediate neighborhood of any chosen pair of parents is very large, and a very large fraction of these offspring will not constitute improvement over the parents. Therefore, even an increased number of mating of good parents is still insufficient to provide a good chance of finding good offspring. For example, Nordin, et al. reported that most crossover events in the standard breeding process produce offspring with less than half of the fitness of their parents [4,5].

One approach to overcoming this problem is to increase the chance of generating improved offspring from a pair of parents by using a customized constructive crossover operator that avoids generating worse offspring altogether. However, designing such operators can be difficult and is likely to be very domain dependent. An open question is whether this approach would result in more effective GP — it is possible that reducing the stochastic element of GP in this way may result in premature convergence or other undesirable restrictions on the GP search. An alternative, simpler, and domain independent approach is to integrate variants of local search techniques into the breeding process to search for good offspring. This can achieve the same effect as constructive crossover operators, though at the cost of a possibly expensive local search. We use this approach to explore the consequences for GP of using constructive crossover operators, and also explore the technique in its own right as modification of the standard GP search.

2 Related Work

Tackett [6] deliberately increased the amount of genetic material used from a selected individual regardless of the frequency of the individual being selected during the

process of generating the next generation. He designed a brood recombination operator. The operator is inspired by the fact that animal species produce far more offspring than are expected to live. It randomly applies crossover N times to two chosen programs to produce $2N$ offspring. After evaluating all offspring, it puts the best two into the next generation and discards rest of the offspring. The brood recombination operator can be categorized as a partial local search operator because it looks for the best state in available states but only looks at $2N$ possible successor states.

Purohit [7] present the concept of FEDS (Fitness, Elitism, Depth limit & Size) crossover. In which they randomly selects individual from the population for the double tournament. In double tournament method there are two tournaments one based on fitness which produces an initial set of winners and a subsequent tournament that selects a subset of those winners based on depth and size limit. Then the best two parent individuals of the tournament are chosen for the crossover operation. From first parent, a subtree is selected and placed at two different positions in the second parent to generate two children. Similarly, from second parent a subtree is selected and placed in the first parent. In this way four individuals are generated from two parents. Then they calculate the fitness, elitism, depth limit and size of the generated children and the two children which have the best result are transferred to the next generation and if the children does not have the better fitness than the parent/child will be retained to the next generation with a 0.5 probability. The FEDS crossover operator can be categorized as a partial local search operator because it looks for the best state in available states but only looks at four possible successor states.

Lang [8] introduced a headless chicken crossover operator which is applied to a chosen program P and a newly (and randomly) generated program R . The operator repeatedly produces offspring from P by replacing a sub-tree of P with a replaced subtree from R until it finds an offspring with greater or equal fitness (problem solving quality) to P 's. The headless chicken crossover operator can be categorized as a hill-climbing local search. It is only a partial local search because it randomly looks for a state better than or equal to the current state and stops once it finds such a state rather than looking at all possible successors. Note that Lang's method is really a mutation rather than crossover, since only one "parent" is chosen from the current generation.

In this paper, we propose to improve the crossover operation, by using the best first technique to generate the offsprings. By using the best first search technique, we remove the randomness of crossover operation. Now we don't have to search for crossover point to generate offsprings.

The remainder of this paper is structured as follows: Section III describes the proposed work. Section IV presents and analyzes the experimental results. Section V draws conclusions.

3 Proposed Work

3.1 Proposed Crossover

In proposed crossover we applied a best first search to improve the overall crossover operation. In this crossover operation, we generate all the possible children from

the parent. Then, we check the fitness of all the generated children. The two children with the highest fitness are chosen. Now we apply the elitism, fitness of the chosen children is compared with the parent. If the fitness of the chosen children is better than the parent, children are transferred to the next generation. If we don't find the children better than the parent, then we transfer the parent to the next generation. Our proposed method is better than previous method in finding solution, because our method has removed the randomness from the crossover operation. There is no need to find the crossover point for performing the crossover. We are generating all the possible combination and finding the best offspring among them. This ensures that, we transfer the optimal offspring to the next generation in every stage. The Proposed method is categorized as the best first search [11], this ensures that it finds the optimal and complete solutions.

3.2 Algorithm for Proposed Crossover

1. Select the best two individuals from the population on the basis of fitness and call them P_1 and P_2 .
2. Swap the sub trees of P_1 and P_2 at different positions and generate all the possible childrens say $ch_1, ch_2 \dots ch_n$.
3. Now select the best 2 children from ch_1, ch_2, \dots, ch_n on the basis of fitness, the children with the highest fitness are chosen (ch_1 and ch_3 , say).
4. Now we apply the elitism.
if $ft(P_1 \text{ and } P_2) < ft(ch_1 \text{ and } ch_3)$ then
transfer ch_1 and ch_3 to the next generation.
5. else
transfer P_1 and P_2 to the next generation.

3.3 Example

As shown in Fig.1, originally, two parent programs P_1 and P_2 are selected. Then, the crossover operation occurs by replacing all the subtree/node of P_1 and P_2 randomly. Generate all the possible combinations C_1, C_2, \dots, C_8 as shown in Fig.1. From all these generated children, we select the best two children on the basis of fitness. The children with the highest fitness are selected. Then we apply the elitism, we compare the fitness of the parent with children. If children fitness is better than parent, children is transferred to the next generation, otherwise parent is transferred.

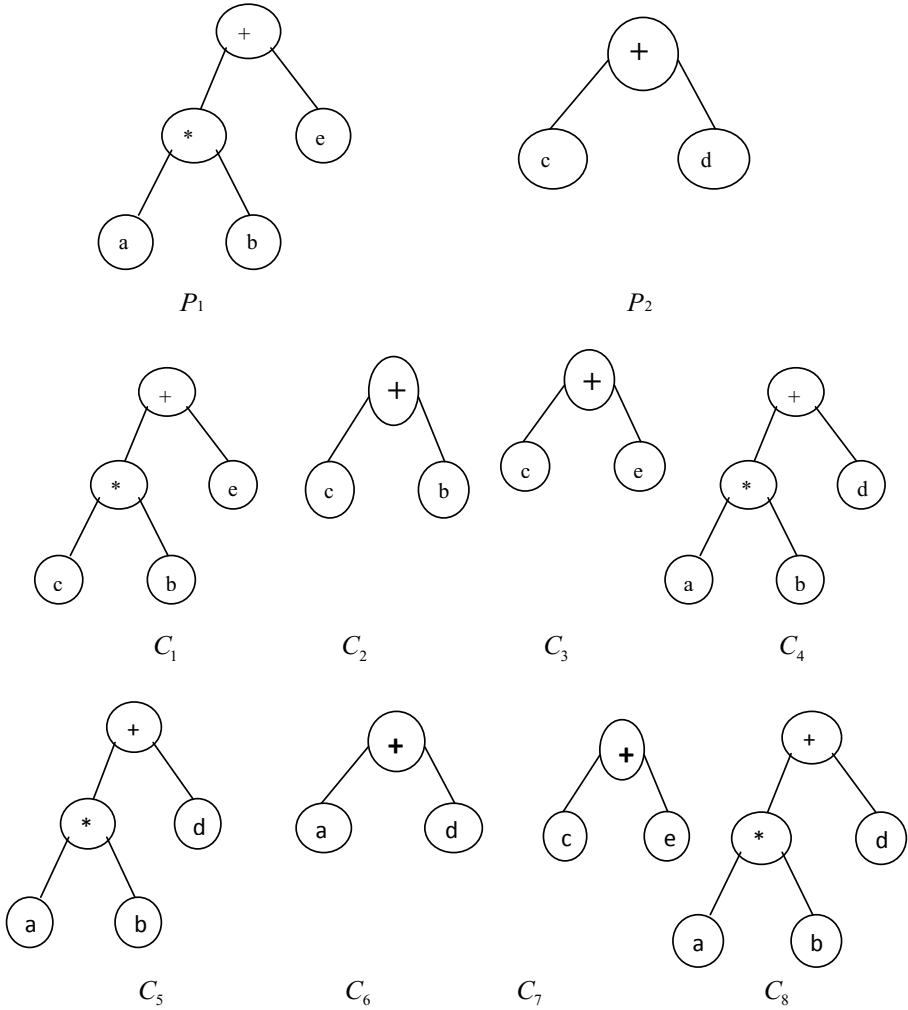


Fig. 1. Proposed Crossover

4 Experimental Results

In this section, proposed GP based Multiclass Classifier is tested. We have used Java 6.0 as a front end tool and Oracle 10g as a back end tool to implement the above proposed method. We tested our proposed method on five benchmark data sets IRIS, WBC, WINE, VEHICLE and BUPA. Table 1 gives a brief description about all the data sets used. All the datasets are taken from the UCI Learning repository [9].

4.1 Data Sets

Four datasets are described in this section.

1) IRIS Data: This is the well-known Anderson's Iris dataset. It contains a set of 150 measurements in four dimensions taken on Iris flowers of three different species or classes. The classes are Iris Setosa, Iris Versicolour and Iris Virginica. The four features are sepal length, sepal width, petal length, and petal width. The data set contains 50 instances of each of the three classes.

2) Wisconsin Breast Cancer (WBC): This data set has 683 data points distributed in four classes. Each data point is represented by ten attributes. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

3) BUPA Liver Disorders (BUPA): It consists of 345 data points in six dimensions distributed into two classes on liver disorders. The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.

4) Wine: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

5) Vehicle: This data was originally gathered at the TI in 1986-87 by JP Siebert. It was partially financed by Barr and Stroud Ltd. The original purpose was to find a method of distinguishing 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects. Measures of shape features extracted from example silhouettes of objects to be discriminated were used to generate a classification rule tree by means of computer induction.

Table 1. Datasets

Name of Data Set	No. of classes	No. of Features
IRIS	3	4
WINE	3	13
WBC	2	9
BUPA	2	6
VEHICLE	4	18

4.2 Parameters

In the approach, we used the tree structure to represent genetic programs [1]. The ramped half-and-half method [10] was used in generating programs in the initial population. The proposed crossover operator is used which were described in Section 3. The point mutation technique is used. The reproduction operator simply copied the best individuals into the population in the next generation to make sure the best individual programs are not lost during evolution.

Table 2 describes the common parameters used for all the data sets. These parameter values are primarily chosen based on the heuristic guidelines on the choice of parameters [6] and an empirical search through initial experiments on GP with the standard crossover operator. Then, we checked those parameter values on the new approach and found that they also could do a reasonably good job.

Table 2. Common Parameters for All Datasets

Parameters	Values
Probability of Crossover Operation	50%
Probability of Reproduction Operation	20%
Probability of Mutation Operation	30%
Population Size	300
Minimum Tree Depth	2
Maximum Tree Depth	6
Number of Generations	40
Double Tournament Size X, Y	9, 5

4.3 Results

For the proposed method, datasets are tested using 10-fold-cross-validation method. Classification accuracy and training time is measured for every datasets as given in Table 3 and Table 4.

We have compared the outcome of our results with HCC method [8], Brood Recombination method [6] and with FEDS crossover method [7] shown in Table 3.

Table 3. Comparison of Conventional Crossover Method, Feds Crossover Method, Hcc Method, Brood Recombination Method With Proposed Crossover Method

NAME OF DATASETS	FEDS crossover Method	HCC Method	Brood Recombination Method	Proposed Crossover Method
	Generalization Accuracy	Generalization Accuracy	Generalization Accuracy	Generalization Accuracy
IRIS	95.42%	95.12%	94.23%	97.67%
WBC	85.56%	86.14%	84.67%	88.65%
BUPA	66.89%	67.32%	68.56%	69.78%
WINE	87.89%	86.47%	86.06%	89.97%
VEHICLE	79.58%	78.25%	77.47%	81.32%

Table 4. GP Run Time

Data Sets / METHOD	IRIS	WBC	BUPA	WINE	VEHICLE
FEDS crossover	0:0:13	0:1:52	0:2:23	0:2:12	0:1:58
HCC	0:0:19	0:2:40	0:2:45	0:2:46	0:2:10
Brood Recombination	0:0:18	0:2:23	0:2:40	0:2:36	0:2:05
Proposed Crossover	0:0:12	0:2:12	0:2:42	0:2:34	0:2:03

It is found that our method outperforms the HCC method, Brood Recombination method and FEDS crossover, point mutation method. Our method improves the accuracy of the classifier with a fair amount. The conventional crossover and mutation technique can be destructive. Rather than generating a good tree with better fitness, the conventional methods can produce the trees with lesser fitness than parents.

In this proposed method we apply the elitism technique which will compare the parent fitness with the child and if the fitness of the parent is better than the child then we transfer the parent to the next generation.

It is also clear from Table 4, that our method will take similar amount of time to find the best solution. Though, it seems to be time consuming, because we generate all the possible combination, when we perform the crossover. But our result shows that we will get the required solution in quicker amount of time. Because by using the best first search technique, always the best path is chosen to reach the goal. By using the best first search in crossover operation, we remove the randomness of crossover operation and also find the optimal solution.

5 Conclusion

In this paper, we proposed a more intelligent crossover operator than the standard crossover operator using best first search in GP. The goal is successfully achieved by developing a new proposed crossover operator. In this approach, a best first search is used to improve the crossover operator. This approach is examined and compared with HCC method, Brood Recombination method and FEDS crossover, point mutation method on five different benchmark datasets. By using the best first always the best path is selected to provide the optimal and complete solution. The results suggest that this new approach outperformed the HCC method, Brood Recombination method and FEDS crossover, point mutation method in terms of the classification accuracy.

References

1. Koza, J.R.: Genetic Programming — On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
2. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D.: Genetic Programming - An Introduction. Morgan Kaufmann, San Mateo (1998)
3. Blickle, T., Thiele, L.: A Mathematical Analysis of Tournament Selection. In: Proceedings of the Sixth International Conference on Genetic Algorithms, pp. 9–16 (1995)
4. Nordin, P., Banzhaf, W.: Complexity compression and evolution. In: Eshelman, L. (ed.) Genetic Algorithms: Proceedings of the Sixth International Conference, Pittsburgh, PA, USA, July 15-19, pp. 310–317. Morgan Kaufmann (1995)
5. Nordin, P., Francone, F., Banzhaf, W.: Explicitly Defined Introns and Destructive Crossover in Genetic Programming. In: Rosca, J.P. (ed.) Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications, pp. 6–22 (1995)
6. Tackett, W.A.: Recombination, Selection, and the Genetic Construction of Computer Programs. PhD thesis, University of Southern California, Los Angeles, CA, USA (1994)
7. Purohit, A., Bhardwaj, A., Tiwari, A., Choudhari, N.S.: Removing Code Bloating in Crossover Operation in Genetic Programming. In: IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, June 3-5 (2011); 978-1-4577-0590-8/11/\$26.00 ©2011 IEEE MIT, Anna University, Chennai
8. Lang, K.J.: Hill Climbing Beats Genetic Search on a Boolean Circuit Synthesis of Koza's. In: Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann (1995)

9. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, Department of Information and Computer Science, Irvine, CA (2007), <http://www.ics.uci.edu/~mllearn/MLRepository.html>
10. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D.: Genetic Programming: An Introduction. On the Automatic Evolution of Computer Programs and Its Application. Morgan Kaufmann, San Mateo (1998)
11. Rich, E., Knight, K., Nair, S.B.: Artificial Intelligence. Tata Mc-Graw-Hill (2009) ISBN: 0070087709

Forecasting Method of Stock Price Based on Polynomial Smooth Twin Support Vector Regression

Shifei Ding^{1,2,3}, Huajuan Huang^{1,2,3}, and Ru Nie^{1,2}

¹ School of Computer Science and Technology,
China University of Mining and Technology, Xuzhou, 221116
dingsf@cumt.edu.cn

² Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Science, Beijing, 100190

³ Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning, 530006

Abstract. The stock price prediction has become an important research topic in Economics. However, the traditional forecasting methods only can be used in linear system, whose prediction accuracy is not satisfactory. In this paper, a new forecasting method of stock price based on polynomial smooth twin support vector regression is proposed. In the proposed method, we firstly construct the polynomial smooth twin support vector regression (PSTSVR) model and prove its global convergence. Then PSTSVR is used as the opening price of stock prediction model. The experimental results on the stock data from the great wisdom stock software show that the proposed method can obtain the better regression performance compared with SVR and twin support vector regression (TSVR).

Keywords: stock price, prediction, polynomial function, smooth, twin support vector regression.

1 Introduction

With the rapid development of the securities industry, stock investment has become an important way of financial management for many families [1-3]. As we know, the stock market has the characteristics of high risk and high profit. Therefore, stock price forecasting has become an important research topic in economics. Up to now, many stock price forecasting methods have been presented [4-5]. Especially, in view of its perfect generalization capability, support vector regression (SVR) has been successfully applied to the field of stock price prediction [6-7]. SVR can well deal with the nonlinear problems introducing the kernel function. So it can obtain more satisfactory prediction value than the traditional methods.

Although SVR owns better generalization compared with other regression methods, its training cost is expensive, i.e., $O(l^3)$, where l is the total size of training data. In order to improve the computational speed, in 2010, Peng [8] introduced a new nonparallel plane regression, termed as twin support vector regression (TSVR). TSVR only need to solve a pair of smaller quadratic programming problems (QPPs), instead

of solving the large one in SVR, which makes TSVR work faster than SVR [9]. Similar to SVR, TSVR solves the QPPs in the dual space. However, this solving method will be affected by time and memory constraints when dealing with the large datasets. In 2010, Chen et al. [10] used the Sigmoid function to approach the objective function of TSVR and then proposed the smooth twin support vector regression (STSVR). STSVR solved the QPPs in original space instead of the dual space. Experimental results showed that STSVR had better regression performance than TSVR. In order to further improve the regression performance of STSVR, looking for a new smooth function with better performance is the key problem.

In this paper, we use polynomial function to approach the non-differentiable term of TSVR and then use the Newton-Armijo method to solve the new model. Based on the above idea, a new smooth twin support vector regression version, termed polynomial smooth twin support vector regression (PSTSVR), is proposed. The experimental results on artificial dataset show that PSTSVR surpasses STSVR, TSVR and SVR.

In view of the perfect regression performance of PSTSVR, we will use PSTSVR to establish the opening price of stock prediction model. The experimental results on the stock data from the great wisdom stock software show that PSTSVR can obtain the better regression performance compared with STSVR, TSVR and SVR.

The paper is organized as follows: In section 2, we propose the PSTSVR model and prove its global convergence. Section 3 deals with experimental results, while section 4 is devoted to concluding remarks.

2 Polynomial Smooth Twin Support Vector Regression

2.1 Twin Support Vector Regression

Let $A \in R^{l \times n}$ denote the input sample matrix, whose row vectors $A_i = (A_{i1}, A_{i2}, \dots, A_{in})$, $i = 1, 2, \dots, l$ are the training samples. Also let $Y = (y_1, y_2, \dots, y_l)^T$ denote the output vector, in which $y_i, i = 1, 2, \dots, l$ are the corresponding response values. TSVR would generate two nonparallel functions around the data points.

For the linear case, TSVR aims at finding a pair of nonparallel functions

$$f_1(x) = w_1^T x + b_1 \quad (1)$$

$$f_2(x) = w_2^T x + b_2, \quad (2)$$

such that each function determines the \mathcal{E} -insensitive down- or up- bounds regressor. The two functions are obtained by solving the following QPPs:

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y - e\mathcal{E}_1 - (Aw_1 + eb_1)\|^2 + C_1 e^T \xi \\ \text{s.t.} \quad & Y - (Aw_1 + eb_1) \geq e\mathcal{E}_1 - \xi, \quad \xi \geq 0 \end{aligned} \quad (3)$$

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y + e\mathcal{E}_2 - (Aw_2 + eb_2)\|^2 + C_2 e^T \eta \\ \text{s.t.} \quad & (Aw_2 + eb_2) - Y \geq e\mathcal{E}_2 - \eta, \quad \eta \geq 0 \end{aligned} \quad (4)$$

where, $C_1, C_2 > 0$; $\mathcal{E}_1, \mathcal{E}_2 > 0$ are the parameters, ξ, η are the slack vectors and e is the vector of ones of appropriate dimensions.

In TSVR, generally, we solve the QPPs in dual space. However, this solving method will be affected by time and memory constraints when dealing with the big datasets. In order to improve the computational speed, TSVR model represented by (3) and (4) would be transformed into two non-smooth optimization problems without constraint by using the plus function and been directly solved in the original space in this section.

According to the KKT theorem, we can get

$$\xi = \max(0, e\mathcal{E}_1 - Y + Aw_1 + eb_1) \quad (5)$$

$$\eta = \max(0, e\mathcal{E}_2 + Y - (Aw_2 + eb_2)) \quad (6)$$

The optimization problems (3) and (4) can be rewritten as

$$\min \quad \frac{1}{2} \|Y - e\mathcal{E}_1 - (Aw_1 + eb_1)\|^2 + C_1 e^T \max(0, e\mathcal{E}_1 - Y + Aw_1 + eb_1) \quad (7)$$

$$\min \quad \frac{1}{2} \|Y + e\mathcal{E}_2 - (Aw_2 + eb_2)\|^2 + C_2 e^T \max(0, e\mathcal{E}_2 + Y - (Aw_2 + eb_2)) \quad (8)$$

Let

$$\phi_1(w_1, b_1) = \max(0, e\mathcal{E}_1 - Y + Aw_1 + eb_1) = (u_1)_+ \quad (9)$$

$$\phi_2(w_2, b_2) = \max(0, e\mathcal{E}_2 + Y - (Aw_2 + eb_2)) = (u_2)_+ \quad (10)$$

where $(u_1)_+$ and $(u_2)_+$ are the plus functions. Apparently, the objective functions of the unconstrained optimization problems (9) and (10) are convex and non-smooth. Therefore, we can't use the traditional numerical optimization method to solve (9) and (10). In order to solve this problem, we will use the polynomial smooth function to approach (9) and (10).

2.2 Polynomial Smooth Function

Weierstrass Theorem [11]. Set arbitrary continuous function $f(x)$, $x \in [m, n]$, then there exists polynomial $P_n(x)$, can make $\lim_{n \rightarrow \infty} \max_{m \leq x \leq n} |f(x) - P_n(x)| = 0$.

The weierstrass theorem shows that any continuous real-valued function in closed interval can be arbitrary approached by the polynomial function. So we can use the polynomial function to approach (9) and (10). In this paper, we will give the common formula of the polynomial smooth function by transforming it to an equivalent infinite series.

Lemma 1 [12]. The two expansion of $m = \frac{1}{2}$ can be expressed as

$$\begin{aligned} \sqrt{1+x} &= 1 + \frac{1}{2}x - \frac{1}{2 \cdot 4}x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^3 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 8}x^4 + \dots = \\ 1 + \frac{1}{2}x - \sum_{n=2}^{\infty} \frac{(2n-3)!!}{(2n)!!}(-x)^n, \quad -1 \leq x \leq 1 \end{aligned} \tag{11}$$

Theorem 1. The plus function u_+ can be transformed to an equivalent infinite series in $[-\frac{1}{k}, \frac{1}{k}]$ as follows.

$$u_+ = \frac{1}{2k} \left(\frac{1+k^2x^2}{2} - \sum_{n=2}^{\infty} \frac{(2n-3)!!}{(2n)!!} (1-k^2x^2)^n \right) + \frac{x}{2} \tag{12}$$

Proof. According to the definition of u_+ , we can get

$$u_+ = \max(0, x) = \frac{|x|+x}{2} = \frac{|kx|}{2k} + \frac{x}{2} = \frac{1}{2k} \sqrt{1+(k^2x^2)} - 1 + \frac{x}{2} \tag{13}$$

According to lemma 1 and (13), u_+ can be rewritten as

$$u_+ = \frac{1}{2k} \left(\frac{1+k^2x^2}{2} - \sum_{n=2}^{\infty} \frac{(2n-3)!!}{(2n)!!} (1-k^2x^2)^n \right) + \frac{x}{2} \tag{14}$$

End.

Theorem 2. The polynomial approximation function for u_+ in $[-\frac{1}{k}, \frac{1}{k}]$ is

$$P_n(x, k) = \begin{cases} x, & x \geq \frac{1}{k} \\ \frac{1}{2k} \left(\frac{1+k^2x^2}{2} - \sum_{l=2}^n \frac{(2l-3)!!}{(2l)!!} (1-k^2x^2)^l \right) + \frac{x}{2}, & |x| < \frac{1}{k}, k > 0 \\ 0, & x \leq -\frac{1}{k} \end{cases} \tag{15}$$

where n is a positive integer.

Theorem 3. $P_n(x, k)$ is defined as (15), it has some characteristics as follows.

- (1) $P_n(x, k)$ has n -order smoothness about x .
- (2) $\lim_{n \rightarrow \infty} \max(P_n(x, k) - x_+) = 0$.

Proof. (1) If $P_n(x, k)$ has n -order smoothness about x , it must meet the following conditions.

$$\begin{aligned}
 P_n\left(\frac{1}{k}, k\right) &= \frac{1}{k}, \quad P_n\left(-\frac{1}{k}, k\right) = 0; \\
 \nabla P_n\left(\frac{1}{k}, k\right) &= 1, \quad \nabla P_n\left(-\frac{1}{k}, k\right) = 0; \\
 \nabla^n P_n\left(\frac{1}{k}, k\right) &= 0, \quad \nabla^n P_n\left(-\frac{1}{k}, k\right) = 0, \quad n \geq 2
 \end{aligned}$$

According to (12), it can be got

$$P_n\left(\frac{1}{k}, k\right) = \frac{1}{k}, \quad P_n\left(-\frac{1}{k}, k\right) = 0$$

We find the partial derivative of x , it can be got

When $n \geq 1$,

$$\nabla P_n(x, k) = \begin{cases} 1, & x \geq \frac{1}{k} \\ \frac{kx}{2} \left(1 + \sum_{i=2}^n \frac{(2l-3)!!}{(2l-2)!!} (1-k^2x^2)^{l-1}\right) + \frac{1}{2}, & |x| < \frac{1}{k}, k > 0 \\ 0, & x \leq -\frac{1}{k} \end{cases}$$

When $n \geq 2$,

$$\nabla^2 P_n(x, k) = \begin{cases} 0, & x \geq \frac{1}{k} \\ \frac{k}{2} \left(1 + \sum_{i=2}^n \frac{(2l-3)!!}{(2l-2)!!} (1-k^2x^2)^{l-1}\right) - \frac{k^3x^2}{2} \sum_{i=2}^n \frac{(2l-3)!!}{(2l-4)!!} (1-k^2x^2)^{l-2}, & |x| < \frac{1}{k}, k > 0 \\ 0, & x \leq -\frac{1}{k} \end{cases}$$

Obviously, $\nabla P_n(x, k)$, $\nabla^2 P_n(x, k)$ and $\nabla^n P_n(x, k)$ ($n > 2$) are existence and continuation in $x = \pm \frac{1}{k}$. So $P_n(x, k)$ has n -order smoothness about x .

(2) According to Weierstrass Theorem, it can be got easily

$$\lim_{n \rightarrow \infty} \max(P_n(x, k) - x_+) = 0.$$

End.

Theorem 3 shows that the polynomial smooth function transformed to an equivalent infinite series can achieve arbitrary precision to approach the plus function when n is large enough.

Because $P_n(x, k)$ has n -order smoothness when $n \geq 2$, Newton-Armijo optimization algorithm can be used to solve the optimization problems without constraint.

If the previous conclusions are extended to nonlinear PSTSVR, the previous conclusions and theorems are also applicable to the nonlinear PSTSVR model.

3 Numerical Experiments and Analysis

In this section, we will do two experiments. To check the performance of PSTSVR, we compare it with SVR, TSVR and STSVR on one group of artificial dataset and in the first experiment. In the second experiment, PSTSVR is used as the opening price of stock prediction model to deal with the stock data from the great wisdom stock software. All the regression algorithms are implemented in MATLAB 7.11 (R2010b) environment on Windows XP running on a PC with 1GB of RAM. In this paper, we only consider the nonlinear case with the Gaussian kernel for all datasets. In all our experiments, we set $C_1 = C_2$ and $\varepsilon_1 = \varepsilon_2$ in the TSVR, STSVR and PSTSVR to degrade the computational complexity of parameter selection. The parameters for all algorithms are selected over the range $\{2^i \mid i = -7, \dots, 7\}$. We evaluated the regression accuracy using 10-fold cross-validation.

3.1 The First Experiment

$\sin ex(x)$ function is usually used to test the performance of the regression methods.

The formula of $\sin ex(x)$ function is $y = \sin ex(x) = e^{-\frac{x}{2.5}} \sin 3x$, $x \in [0, 4\pi]$.

To effectively check the performance of PSTSVR, training data points are perturbed by some different types of noises, which include the Gaussian noises and the uniformly distributed noises. After adding four types of noises into training data points, $\sin ex(x)$ is expressed as $y_i = \sin ex(x_i) + \xi_i$, where (a) $\xi_i \in N(0, 0.1^2)$, (b) $\xi_i \in N(0, 0.2^2)$, (c) $\xi_i \in U(-0.2, 0.2)$ and (d) $\xi_i \in U(-0.4, 0.4)$.

$U[a, b]$ represents the uniformly random variable in $[a, b]$ and $N(c, d^2)$ represents the Gaussian random variable with c means and variance d^2 .

In order to improve the reliability of results comparison, for each kind of noise, fifteen independent groups of noisy samples are randomly generated which respectively consists of 500 training samples and 500 test samples. And the test points are uniformly sampled from the $\sin ex(x)$ function without any noise. In order to test the impact on the performance of PSTSVR using different values of the order of polynomial n , we conduct experiments on $\sin ex(x)$. The experimental results are shown as Table 1 when the values of n are set 1, 2, 3, 4, 5 respectively. This experiment including 4 indicators, from top to bottom are SSE and CPU running time.

Table 1. The experiment results of PSTSVR when n take different values

Noise Type	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
	SSE Time(s)	SSE Time(s)	SSE Time(s)	SSE Time(s)	SSE Time(s)
a	0.1307 ± 0.0021 2.1214	0.1302 ± 0.0024 2.1345	0.1202 ± 0.0019 2.1356	0.1202 ± 0.0014 2.2524	0.1202 ± 0.0012 2.2549
	0.1353 ± 0.0062 2.3087	0.1353 ± 0.0058 2.3157	0.1351 ± 0.0062 2.3164	0.1351 ± 0.0058 2.4120	0.1351 ± 0.0055 2.5641
c	0.0959 ± 0.0075 2.3041	0.0956 ± 0.0074 2.3047	0.0952 ± 0.0045 2.3055	0.0952 ± 0.0043 2.3341	0.0952 ± 0.0039 2.3445
	0.1248 ± 0.0056 2.2145	0.1247 ± 0.0045 2.3431	0.1244 ± 0.0021 2.3553	0.1244 ± 0.0019 2.3941	0.1244 ± 0.0016 2.4312

From table 1, we find that with the increase of n , PSTSVR yields the better regression performance, but it costs higher run-times at the same time. The possible reason is that the approximation accuracy of smooth polynomial function will be higher with the increase of n . However, it will cost more run-times because the polynomial function is more complex. From table 1, we also find that regression performance and run-times are the most suitable for PSTSVR when $n = 3$. Therefore, in order to demonstrate the performance of PSTSVR with other algorithms, we conduct the experiments on $\sin ex(x)$ when $n = 3$. The experimental results are shown as Table 2. Figs.1-4 show the one-run simulation results of SVR, TSVR, STSVR and PSTSVR for four types of noise.

Table 2. comparisons of PSTSVR with other algorithms ($n = 3$)

Noise	Algorithms	SSE	SSE/SST	SSR/SST	Time(s)
a	SVR	0.1927 ± 0.0024	0.0067 ± 0.0031	0.9528 ± 0.0325	4.1356
	TSVR	0.1851 ± 0.0021	0.0058 ± 0.0054	0.9534 ± 0.0243	2.3341
	STSVR	0.1554 ± 0.0047	0.0055 ± 0.0035	0.9715 ± 0.0244	2.3242
	PSTSVR	0.1202 ± 0.0019	0.0044 ± 0.0023	0.9823 ± 0.0427	2.1356
b	SVR	0.1952 ± 0.3464	0.0041 ± 0.0032	0.9428 ± 0.0021	5.6784
	TSVR	0.1984 ± 0.5243	0.0049 ± 0.0054	0.9514 ± 0.0031	2.4612
	STSVR	0.1637 ± 0.1471	0.0035 ± 0.0032	0.9684 ± 0.0052	2.3311
	PSTSVR	0.1351 ± 0.0062	0.0018 ± 0.0027	0.9934 ± 0.0011	2.3164
c	SVR	0.1856 ± 0.0074	0.0032 ± 0.0013	0.9489 ± 0.0152	6.1235
	TSVR	0.1135 ± 0.0071	0.0019 ± 0.0071	0.9674 ± 0.0234	2.7983
	STSVR	0.0993 ± 0.0043	0.0019 ± 0.0023	0.9823 ± 0.0162	2.5263
	PSTSVR	0.0952 ± 0.0045	0.0016 ± 0.0013	0.9875 ± 0.0152	2.3055
d	SVR	0.1848 ± 0.0056	0.0077 ± 0.0025	0.9438 ± 0.0125	6.245
	TSVR	0.1764 ± 0.0078	0.0069 ± 0.0041	0.9524 ± 0.0344	3.0831
	STSVR	0.1245 ± 0.0061	0.0068 ± 0.0012	0.9726 ± 0.0213	2.7413
	PSTSVR	0.1244 ± 0.0021	0.0054 ± 0.0023	0.9845 ± 0.0152	2.3553

From table 2, we can see that PSTSVR obtains the smallest SSE, SSE/SST, and the largest SSR/SST for these four types of noise. Besides, the learning CPU time of PSTSVR is least. This indicates that PSTSVR not only obtains higher learning speed but also has better regression performance. Figs.1-4 show that the fitting capacity of PSTSVR is perfect.

3.2 The Second Experiment

The above experimental results have shown that PSTSVR owns perfect regression performance. So in this section, we would use PSTSVR model as the opening price of stock prediction model. As we know, effectively predicting the opening price can provide powerful information for observing changes of the stock market as a whole, which is meaningful for the stock prediction. In this section, the opening price data of Shanghai composite index are got from the great wisdom stock software (<http://www.gw.com.cn>), which includes 5279 trading days data from 1990.11.19-2012.07.31. The Shanghai composite index data contains 6 indicators, which respectively are the opening price, the highest index value, the minimum index value, close index, trading volume and trading turnover.

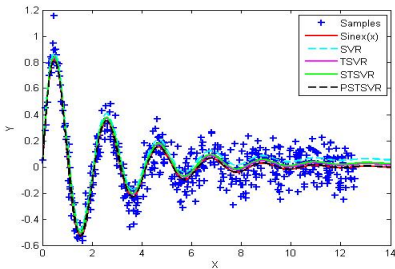


Fig. 1. Predictions of SVR, TSVR, STSVR and PSTSVR on $\sin ex(x)$ function with noise $N(0, 0.1^2)$

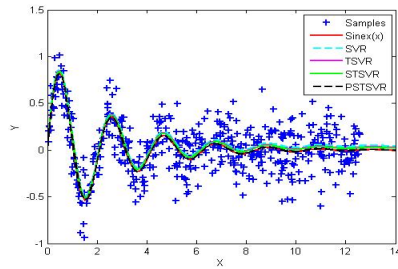


Fig. 2. Predictions of SVR, TSVR, STSVR and PSTSVR on $\sin ex(x)$ function with noise $N(0, 0.2^2)$

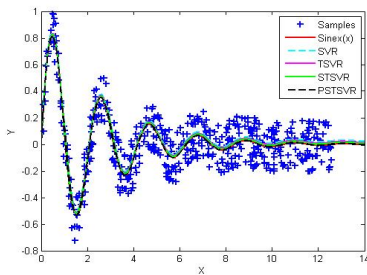


Fig. 3. Predictions of SVR, TSVR, STSVR and PSTSVR on $\sin ex(x)$ function with noise $U(-0.2, 0.2)$

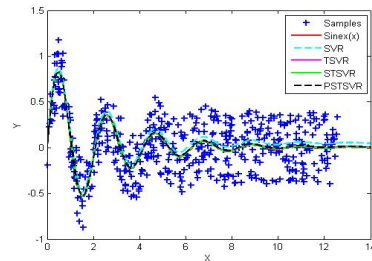


Fig. 4. Predictions of SVR, TSVR, STSVR and PSTSVR on $\sin ex(x)$ function with noise $U(-0.4, 0.4)$

To predict the opening price, we use PSTSVR to establish the regression model. Before do it, we would give some assumptions as follows.

Assume that the opening price data of the day are related to the highest index value, the minimum index value, close index, trading volume and trading turnover of the before day, i.e., the highest index value, the minimum index value, close index, trading volume and trading turnover of the before day are as the independent variables and the opening price of the day is as the dependent variable. Based on the above assumption, we establish the regression model. Similar to the first experiment, to check the performance of PSTSVR, we compare it with SVR, TSVR and STSVR. Table 3 shows the average results of SVR, TSVR, STSVR and PSTSVR with 15 independent runs on the Shanghai composite index data. Fig.5-6 are the fitting results of opening price using STSVR and PSTSVR respectively.

Table 3. Result comparisons of SVR, LS-SVR, TSVR and PLTSVR on stock opening price data

Algorithm	SSE	SSE/SST	SSR/SST	Time(s)
SVR	0.0045 ± 0.176	0.0182 ± 0.0213	0.9829 ± 0.0114	5.2423
TSVR	0.0043 ± 0.185	0.0187 ± 0.0545	0.9821 ± 0.0195	1.1622
STSVR	0.0031 ± 0.183	0.0164 ± 0.0232	0.9904 ± 0.0254	1.1503
PSTSVR	0.0027 ± 0.182	0.0155 ± 0.0123	1.0018 ± 0.0112	1.1343

It can be seen that PSTSVR has lower SSE, SSE/SST values and higher SSR/SST values than other algorithms from table 3. Furthermore, for learning time, PSTSVR is fastest than other three algorithms. It indicates our algorithm is an effective method for predicting the stock opening price.

From fig.5 and fig.6, we can easily find that the fitting result of PSTSVR is better than STSVR on stock opening price dataset.

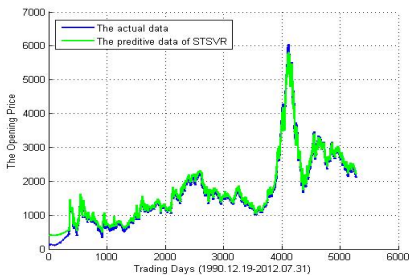


Fig. 5. The fitting results of STSVR on stock opening price

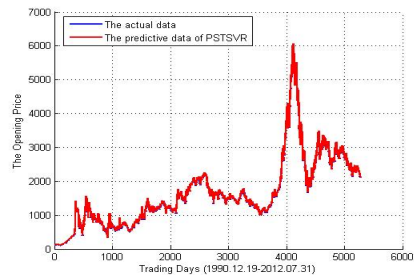


Fig. 6. The fitting results of PSTSVR on stock opening price

4 Conclusions

The volatility of the stock market has gradually become an important index to measure the level of economic development in China. However, the traditional forecasting

methods only can be used in linear system, whose prediction accuracy is not satisfactory. In this paper, a new forecasting method of stock price based on polynomial smooth twin support vector regression is proposed. In the proposed method, we firstly construct the polynomial smooth twin support vector regression (PSTSVR) model and prove its global convergence. Then we use PSTSVR as the opening price of stock prediction model. The experimental results show that the proposed method can obtain the better regression performance compared with SVR and twin support vector regression (TSVR). It is an effective method of prediction of stock price.

Acknowledgments. This work is supported by the National Key Basic Research Program of China (No. 2013CB329502), the National Natural Science Foundation of China (No.41074003), and the Opening Foundation of Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis.

References

1. Zarandi, M.H.F., Zarinbal, M., Ghanbari, N.: A New Fuzzy Functions Model Tuned by Hybridizing Imperialist Competitive Algorithm and Simulated Annealing Application: Stock price Prediction. *Information Sciences* 222, 213–228 (2013)
2. Hsu, C.M.: A hybrid Procedure with Feature Selection for Resolving Stock/Futures Price Forecasting Problems. *Neural Computing & Applications* 22(3), 651–671 (2013)
3. Wei, L.Y.: A GA-weighted ANFIS Model Based on Multiple Stock Market Volatility Causality for TAIEX Forecasting. *Applied Soft Computing* 13(2), 911–920 (2013)
4. Chen, M.Y.: A Hybrid ANFIS Model for Business Failure Prediction Utilizing Particle Swarm Optimization and Subtractive Clustering. *Information Sciences* 220, 180–195 (2013)
5. Kazem, A., Sharifi, E., Hussain, F.K., et al.: Support Vector Regression with Chaos-Based Firefly Algorithm for Stock Market Price Forecasting. *Applied Soft Computing* 13(2), 947–958 (2013)
6. Luo, L.K., Chen, X.: Integrating Piecewise Linear Representation and Weighted Support Vector Machine for Stock Trading Signal Prediction. *Applied Soft Computing* 13(2), 806–816 (2013)
7. Wang, B.H., Huang, H.J., Wang, X.L.: A Support Vector Machine Based MSM Model for Financial Short-Term Volatility Forecasting. *Neural Computing & Applications* 22(1), 21–28 (2013)
8. Peng, X.J.: TSVR: An Efficient Twin Support Vector Machine for Regression. *Neural Networks* 23, 365–372 (2010)
9. Zhong, P., Xu, Y.T., Zhao, Y.H.: Training Twin Support Vector Regression via Linear Programming. *Neural Computer & Application* 21, 399–407 (2012)
10. Chen, X.B., Yang, J., Liang, J.: Smooth Twin Support Vector Regression. *Neural Computer & Application* 21, 505–513 (2013)
11. Liu, Y.Q., Liu, S.Y., Gu, M.G.: Self-training Polynomial Support Smooth Semi-supervised Support Vector Machines. *Journal of System Simulation* 21(18), 5740–5743 (2009)
12. Ren, B., Cheng, L.L.: Polynomial Smoothing Support Vector Regression. *Control Theory & Applications* 28(2), 261–265 (2011)

ApLeafis: An Android-Based Plant Leaf Identification System^{*}

Lin-Hai Ma¹, Zhong-Qiu Zhao^{1,2}, and Jing Wang¹

¹ College of Computer Science and Information Engineering,
Hefei University of Technology, Hefei 230009, China

² Department of Computer Science, Hong Kong Baptist University, Hong Kong, China

Abstract. To automatically identify plant species is very useful for ecologists, amateur botanists, educators, and so on. In this paper, an Android-based mobile application designed to automatically identify plant species by the photographs of tree leaves is described. In this application, one leaf image can be either a digital image from the existing leaf image database or a picture collected by a camera. The picture should be a single leaf placed on a light and untextured background without other clutter. The identification process consists of totally three steps: leaf image segmentation, feature extraction, and species identification. The demo system is evaluated on the ImageCLEF2012 Plant Identification database which contains 126 tree species from the French Mediterranean area. The output of the system to users is the top several species which match the query leaf image the best, as well as the textual descriptions and additional images about leaves, flowers, etc., of theirs. Our system works well with state-of-the-art identification performance.

1 Introduction

Past 200 years of human activities have led to the extinction of hundreds of species of plants. It is crucial to protect the diversity of plants. The identification of the plant species is the most important step for further protection. However, existing identification methods mainly rely on scientist or specialized knowledge. Thus, an automatic plant species identification system is desired. With the consideration of geometric distribution of plants, the identification system based on mobile devices has attracted a lot of attention. Kumar et al. [1] designed the first mobile application for plant species identification, namely Leafsnap. However, Leafsnap could perform the identification only when accessed to the Internet and the query image must be taken by the local camera. Besides, Leafsnap is based on IOS, while Android is more popular mobile operation system than IOS. Thus, we design the first Android-based plant image retrieval system based on content-based image retrieval (CBIR) technique.

^{*} This research was supported by the National Natural Science Foundation of China (No. 61005007), the 973 Program of China (No. 2013CB329604), the 863 Program of China (No. 2012AA011005), and the Hong Kong Scholars Program (No. XJ2012012).

In this paper, we will introduce the major steps of plant identification in our system described in Fig. 1. Given a query image taken by the camera or already existed in the local database, we first segment it to binary image by threshold segmentation, and move the stem through Tophat (Section 2). Then we extract the features of PHOG, color, etc., and fuse them to form the final feature space (Section 3). The histogram intersection is applied to predict the class (Section 4). The simulation results are listed in Section 5, and finally the basic functions of the system are shown in Section 6. The local database in our system is “ImageCLEF” [2] which has a coverage of the 126 tree species in French. The application works well on any device with Android system.

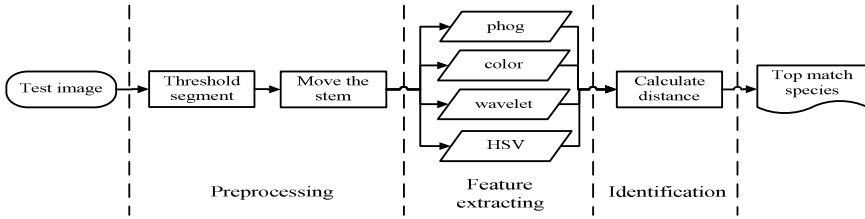


Fig. 1. General scheme of plant identification by leaves

2 Leaf Image Preprocessing

In our system, the input image must be against light, with a untextured background. Considering both the software running speed and the segmentation results, we adopt the threshold segmentation based on color pixel [3] then remove the small bad regions and the stem by Tophat. The details are given as follows.

Threshold Segmentation: It is an efficient technique for image segmentation. First, we convert the RGB space into gray, and then binarize the image according to the gray level, defined as:

$$b(x, y) = \begin{cases} 1, & \text{if } g(x, y) > M \\ 0, & \text{if } g(x, y) \leq M \end{cases} \quad (1)$$

where $g(x, y)$ are points of the gray level image pixels, $b(x, y)$ are points of binary level image, M is the mean value of the image matrix.

Removing the Stem: By human observation, there is no uniform layout, width or length for leaf stems, which usually results in recognition failures. Therefore, after filtering to eliminate the influence of noises, we remove the stem by Tophat [4]. Tophat transformation can determine the thin structures by designing a structuring disc element with a diameter longer than the width of stem. For a binary image B and the structuring element S , the Tophat process is: $B - B \oplus S$, where \oplus denotes the operation of an erosion followed by a dilation.

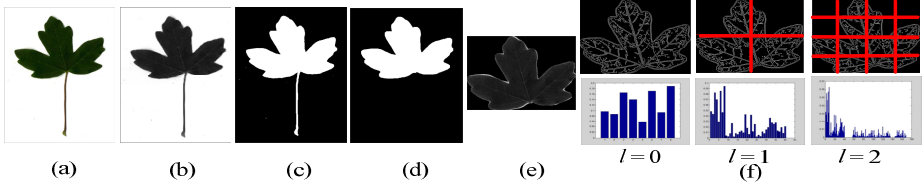


Fig. 2. We turn an original RGB image (a) into gray (b), obtain the binary image (c) by the threshold technique, remove the stem (d) by the tophat transformation, and then crop it (e). Shape spatial pyramid representation (f). Top row: a leaf image and grids for levels $l = 0$ to $l = 2$; Below: histogram representations corresponding to each level.

After removing the stem, we find that the result is better if we just extract features from the region of leaf. So we crop it at the end of the segmentation (Fig. 2).

3 Feature Extraction

The algorithms of feature extraction should be as simple as possible to meet the requirements of real-time, so our algorithm adopts PHOG [5], color [7], HSV and wavelet [8] features to find the best combination of feature space.

Pyramid Histograms of Oriented Gradients (PHOG): We first use edge detector to identify the edge contour of the leaf and divide the leaf spatial pyramid, as shown in Fig. 2(f). Then we use the idea of histogram of oriented gradient (HOG) [6]. This process is recursively repeated to level l , and the grid at level l has 2^l cells along each dimension. The final PHOG descriptor for an image is a concatenation of all vectors at each pyramid resolution, and thereby contains the spatial information of the objects in the image. If $sobel_x$, $sobel_y$ are the two Sobel operators, the gradients at the point (x, y) of image I can be defined as $I \times sobel_x$ and $I \times sobel_y$, then the gradient amplitude $|I|$ and the orientation of the gradient θ at the point (x, y) is: $|I| = \sqrt{I_x^2 + I_y^2}$, $\theta = \arctan(I_x/I_y)$. In our system, we set $N = 8$ and $L = 2, 3$ respectively, and PHOG features are extracted to represent their local shape and spatial layouts.

Color Features: Stricker [7] calculated three central moments (Mean E_i , Standard deviation σ_i and Skewness s_i) for every channel in an image, defined as:

$$E_i = \sum_{j=1}^N \frac{1}{N} p_{ij} \cdot \sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^N (p_{ij} - E_i)^2\right)}, s_i = \sqrt[3]{\left(\frac{1}{N} \sum_{j=1}^N (p_{ij} - E_i)^3\right)} \quad (2)$$

where p_{ij} is the j -th image pixel at the i -th channel. Here, we get an 81-dimensional vector (3 moments for each 3 color channels of 3×3 grids) as color feature vector.

HSV Features: Based on the substantial analysis, we convert the RGB image to HSV image, and divide H into 32 parts, S and V into 8 parts separately. The HSV descriptor is with 48 dimensions.

Wavelet Features [8]: The simple Haar Transform has been used here to determine the scaling and wavelet functions: $\psi^H(x, y) = \varphi(y)\psi(x)$, $\psi^V(x, y) = \varphi(x)\psi(y)$, $\psi^D(x, y) = \varphi(x)\psi(y)$ where $\psi^H(x, y)$, $\psi^V(x, y)$, $\psi^D(x, y)$ measures variations along horizontal, vertical, diagonals respectively. The translated basis function is defined as: $\psi^i_{j,m,n}(x, y) = 2^{j/2}\psi^i(2^j x - m, 2^j y - n)$, $i = \{H, V, D\}$ and the discrete wavelet transform of function $f(x, y)$ of size M by N is formulated as:

$$w^i_{\psi}(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \psi^i_{j,m,n}(x, y) \quad (3)$$

where $w^i_{\psi}(j, m, n)$ represents the horizontal, vertical and diagonal details. Here we divide the input image to 3×3 grids, and use the coefficients of three directions on three scales. Finally we obtain the wavelet descriptor with 81 dimensions.

Feature Selection: For a leaf, the color changes with the season, but the shape remains unchanged. Thus, the shape-based features are more suitable here. We evaluate individual features by the Nearest Neighbors method to observe their relative merits, and the result can be found in Section 5. Color features are also evaluated but not discriminative according to our experimental results, so they do not appear in the final feature list.

4 Plant Species Identification

After extracting features, we need to calculate the distance automatically. As the used features are all based on the statistical theory, the histogram intersection may work better than the usual Euclidean distance as the metric [9]. In this paper, the comparison experiments were made between them, and showed that the histogram intersection works better. The definition of Histogram intersection is: $d(a, b) = 1 - \sum_i^N \min(a_i, b_i)$,

where N is the feature dimensionality. After removing the duplicate categories, the list of top matched species are shown to users.

5 Experimental Results

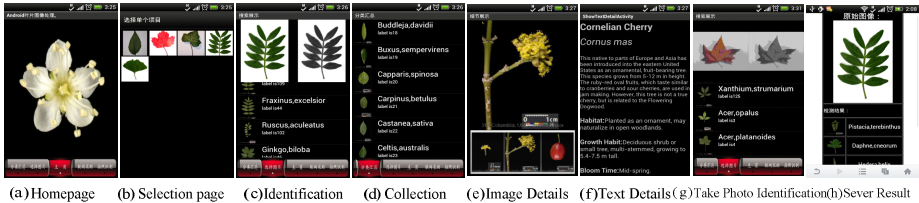
The database used in this system comes from the ImageCLEF2012 competition, and it contains 126 tree species in the French Mediterranean area. The database consists of three kinds of images: scans, scan-like photos and unrestricted photos. We extract features from 4870 scans and scan-like photos and use them as the search database in the current application version. However, due to the constraints of the mobile phone, a series of cross-validation experiments are taken on Matlab in order to determine the most suitable descriptors for identification. We choose 2500 images for training and 2370 for testing. The results are shown in Table 1.

Table 1. The descriptors along with their identification accuracies

ID	Feature	Level	Size	Accuracy	
				histogram intersection	Euclidean distance
f1	Color	---	81	0.68±0.011	0.63±0.00
f2	HSV	---	48	0.65±0.011	0.59±0.00
f3	Wavelet	---	81	0.62±0.006	0.53±0.00
f4	Phog	L=2	40	0.68±0.010	0.66±0.01
		L=3	168	0.78±0.010	0.75±0.01
f5	Phog+tophat+cropped	L=2	40	0.71±0.005	0.69±0.00
		L=3	168	0.81±0.008	0.76±0.00
f2,f3,f5	---	L=2	169	0.89±0.004	0.80±0.00
		L=3	297	0.90±0.005	0.79±0.00

6 The Software System

In this work, we introduce our Android application for automatic plant identification. The application is written in Java language, and it contains several parts as shown in Fig. 3: **The homepage** (a), with several beautiful pictures of plants to display cycling for users every a few seconds; **The selection page** (b), where users can click the second button to open the picture library, and choose one image to identify. Then the system will show the list of matched species for users to make the final judge themselves (c), and users can compare the actual plant with the high-quality photographs (e, f) if they check one of the items; **The collection page** (d), with the list of 126 plants for users, where users can click on the search button to look for some special plants by the name. Similar to the selection page, users will get the details if they check it; **The identification on server page** (h): considering the limitations of the mobile phone, users can use a server to perform some complex calculations. After uploading the image and checking the connection button, a more accurate result will be obtained from the server and the real-time requirements can also be met as well; **The camera page** (f), where users can use the camera to take picture and identify it if they meet a real leaf. In this case, the picture should match the requirements: a single leaf placed on a light and untextured background without other clutter; otherwise, it may cause the failure of identification.

**Fig. 3.** The Mobile Phone Screenshots

Compared with the existing software based on the Apple device [1], our application works on a more popular operate system – Android, with more powerful functions such as checking the details, identifying the pictures of leaves, and so on. Besides, users can identify the species without Internet, and select digital images as query image. The last but the least advantage is that our system is more quickly which is more attractive in practice.

7 Future Work

Our system relies on many computer vision techniques. In the future work, we will promote our system by improving the segmentation method and exploring more accurate feature extraction methods.

References

1. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B.: Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 502–516. Springer, Heidelberg (2012)
2. Zheng, P., Zhao, Z.Q., Herve, G.: ZhaoHFUT at ImageCLEF 2012 Plant Identification Task (2012)
3. Haralick, R.M., Shapiro, L.G.: Image Segmentation Techniques. *Computer Vision, Graphics, and Image Processing* 29(1), 100–132 (1985)
4. Burger, W., Burge, M.J.: *Digital Image Processing* (2008)
5. Chen, J., Bai, Y.: Classification of Smile Expression Using Hybrid PHOG and Gabor Features. In: ICCASM, vol. 12, pp. 417–420 (2010)
6. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR 2005, vol. 1, pp. 886–893 (2005)
7. Stricker, M.A., Orengo, M.: Similarity of Color Images, pp. 381–392. International Society for Optics and Photonics (1995)
8. Burrus, C.S., Gopinath, R.A., Guo, H., et al.: *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice hall, Upper Saddle River (1998)
9. Wang, M., Hua, X.S., Tang, J., Hong, R.: Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE Trans. Multimedia* 11(3), 465–476 (2009)

Microblog Searching Module Based on Community Detection

Pengxiang Lin, Tingting He*, and Yong Zhang

School of Computer, Central China Normal University, WuHan, China
linpx1007@126.com, tthe@mail.ccnu.edu.cn, ychang.cn@gmail.com

Abstract. In this paper, we investigate the current software architecture of Twitter searching, and propose a new Microblog Searching Module (MSM) to retrieve microblog messages. MSM mainly consists of three parts. The first one is community detection with Label-Influence-Algorithm (LIA). We have conducted series of experiments in two data sets downloaded from the Sina-Microblog. And the results show that the modularity measure Q of the communities discovered by LIA is well improved. The second one is extracting microblog tags of a microblog user and the community. The last part is designing a module to expand the query word using the HowNet instead of the exact word match. The application of the Microblog Searching Module proves that the module can search the interesting topic and persons conveniently.

Keywords: Sina-Microblog, social network, Label-Influence-Algorithm, microblog tag, query expansion.

1 Introduction

Nowadays, the popularity of social media has created huge number of web data that has led to an evolution in the way people consume information [1]. Microblog is a typical social media. A large majority of the microblog users update their microblogs very frequently. Our study confirms that there are some microblog users who do seriously follow someone because of common interests instead of just playing a number game [2]. In other words, there are some communities in the microblog network for common interests.

There are about four hundred million users in Sina-Microblog. Users post millions of short messages every second. But the search approach is kept limited to basic text matching. The task of collecting, searching and analyzing microblogs is rather challenging for various reasons [3]. First, it is rather different to search all microblogs using API. Second, because the size of microblog message is limited to 140, this clearly makes any attempt to use natural language processing tools limited. Third, the microblog world is a dynamic network. This would suggest that the microblog searching module must combine the natural language processing with the social network [4].

* Corresponding author.

The main contribution of this paper is listed below. We propose a module named Microblog Searching Module to search the real-time microblogs on the specific field. The module mainly contains three parts. The first one is an efficient algorithm to detect communities in microblog network. The second one is extracting microblog tags of a microblog user and the community. At last, we design a method to expand the query word using the Hownet instead of the exact word match. The application of the module shows that it works well.

2 Related Work

Anderson et al [5] have suggested a software architecture for tweet collection that allows users to search for tweets that match a query word. The system provides the links to Twitter friends and followers. This is built using MVC, Hibernate, JPA in conjunction with Tomcat, MySql and Lucene.

Perera et al [6] have suggested an architecture for tweet collection that makes use of tweet API, Python and MySql. They extract the user ids and the time-stamps of the tweets, and use the collected data to summarize the inter-connection times between tweets and the number of retweets. They get the conclusion that the new tweets of a user can be modeled as a Poisson Process. And the number of retweets can be seen as a geometric distribution.

Connor et al [7] have suggested a search architecture for Twitter named TweetMotif. Through a faceted search interface, TweetMotif groups messages by frequent significant words. The topic extraction system is on the base of syntactic filtering and language modeling. TweetMotif can be used to deflate rumors, uncover scams, and track political protests in real-time.

3 Microblog Searching Module

3.1 Community Detection

On the base of label propagation [8], we propose an efficient algorithm named Label-Influence-Algorithm (LIA) to detect community in microblog network. In the microblog network, there are two kinds of ‘neighborhood’. They are friends and followers. A person tends to in the same community with a more influential one if they are familiar with each other. If there are some people with high influence in a community, the community will be steadier [9].

We can describe the LIA in the following steps.

1. Get the seed set and initial the labels at all nodes in the network. First set a parameter k that means every k nodes choose the most influential one setting a seed labeled i (from 1 to N/k , N is the total number of all microbloggers) in the node set. For a given node x without in the seed set, $C_x(0) = 0$;
2. For each node, if its friends have labels, get these labels and count the number of every label. Choose the maximum number label, and if there are more than one la-

bel in the maximum number, choose the most influential one, not randomly. Previously, a microblog owner's influence is often measured by his node in-degree, for example, the number of followers. However as observed in previous studies [10], in-degree does not accurately capture the notion of influence. Here we propose a formula to measure the 'influence'. That is

$$\text{Influence}(x) = \frac{f(x)}{d(x)} + c(x) + a(x) \quad (1)$$

$f(x)$ stands for the number of user's followers, and $d(x)$ stands for the number of user's friends. $c(x)$ stands for the number of being commented in all microblogs. $a(x)$ stands for the number of one microblogger being @ by others.

In order to measure the performance of community detection, Newman and Girvan [10] had proposed an evaluating method called modularity measure Q . This evaluating method is generally accepted by academia. For the communities without common parts, modularity measure Q is defined as following:

$$Q = \sum_{c \in C} \left[\frac{I_c}{E} - \left(\frac{2I_c + O_c}{2E} \right)^2 \right] \quad (2)$$

I_c stands for the total number of edges that both of the two vertices are in the same community C . O_c stands for the number of edges that only one of the two vertices in the same community C . E stands for the total number of edges in the whole network.

3.2 Extract Microblog Tag

Because the size of microblog message is limited to no more than 140 characters, a user must refine the language to express his meaning [11]. If only use the microblog messages, we cannot get the interest topic well. There are probably many comments for a microblog message, and these comments may contain much information. We choose all the microblog messages the user proposed together with all the comments.

In order to get user's tags, firstly, we have to do Chinese word tokenization and part of speech tagging using the ICTCLAS2013 (free software architecture for Chinese word tokenization) [12]. The noun and verb of the microblog messages can express the user's interests well. So we get a word set including a number of noun and verb words for per user.

In order to extract microblog tags, we choose the TFIDF ranking algorithm [13]. The biggest advantage to this approach is that it is simple but efficient. The TFIDF value of word i from this user's messages is computed as

$$\text{tfidf}_{iu} = \frac{n_{iu}}{\sum_j n_{ju}} \log\left(\frac{U}{U_i}\right) \quad (3)$$

Where n_{iu} is the count of word i in user u 's messages, U_i is the number of users whose messages contain word i , and U is the total number of users in the microblog corpus. For each user, words with top N TFIDF values are selected as his tag.

3.3 Query Expansion

In this module, we can get an expansion set of a query word to search, instead of the single word match. We do the query expansion using the HowNet. In the HowNet, sememe is the basic and inseparable unit and we can use sememes to describe the relationship between concepts and concepts, and also concepts and attributes [14].

There are two cases to do the query expansion using the HowNet [15].

Firstly, we map the query word to series sememes, and those words have the same first-sememe are the expansion of the query word.

Like “做饭” DEF={cook|烹调 : PatientProduct={edible|食物}}, “烹饪” DEF={cook|烹调}, “烹饪” can be the query expansion word of “做饭”.

Secondly, put all the words with the hyponymy relationship to the expansion set because sememes are organized by hierarchical structure and hyponym word is usually explained by hypernym word.

Like “控制” is the hypernym word and “管理”, “处理”, “安排” are the hyponym words. So “管理”, “处理”, “安排” can be the query expansion words of “控制”.

4 Experiment

4.1 Data Set

For there is no common microblog corpus, we employ Sina-Microblog API to download two data sets. The first one is 1201 users and 8486 single-track edges (DS1). And the second data set is 737 users and 2602 bipartite edges (DS2). In order to assure the reliability of the experiment, 737 users are from those have bilateral friends in 1201 users.

4.2 Community Detection Result

We conducted six experiments to test the LIA. The Experiment 1 is using label propagation with DS1. The Experiment 2 is using LIA with DS1. The Experiment 3 is using K-means with DS1. The Experiment 4 is using label propagation with DS2. The Experiment 5 is using LIA with DS2. The last one is using K-means with DS2. In Experiment 3 and Experiment 6, we set the number of communities as 120.

As shown in Table 1, these are the modularity measure Q of six experiments.

It is easy to see that the Q value got by LIA in Experiment 2 with DS1 is higher than Experiment 1 and Experiment 3. The Q value got by LIA in Experiment 5 with DS2 is higher than Experiment 4 and Experiment 6. According to the Q value got by the experiments, we can get the conclusion that the performance of LIA is better than label propagation and K-means.

Table 1. Modularity measure Q

Experiment	Data set	Q
1(label propagation)	DS1	0.453
2(LIA)	DS1	0.459
3(K-means)	DS1	0.182
4(label propagation)	DS2	0.567
5(LIA)	DS2	0.688
6(K-means)	DS2	0.204

It is proved by these results that considering the microblogger's influence and using bipartite relationships are wise choices to detect community in microblog networks. Choosing the label according to the more influential neighbors can highly raise the value of modularity measure Q and reduce the uncertainty of the communities in microblog networks.

Table 2. A community got from the data set with LIA

Label	ID	Nickname	Tag
116	1865631071	王磊_Mir	music searching voice recognition
116	1642011213	yuanchao	search engine programmer
116	1671742992	小暖 _WuLiang	recommendation system opinion mining
116	1401425594	CloudyBI	data analysis data warehouse
116	1639378393	叶子ku	NLP search engine
116	1822006777	AlexPitt	architect PMP

As shown in Table 2, this is a part of the community got with LIA. The first column is the label get by LIA. The second one is microblogger's id, and the third is one's nickname. The last column is the tag that the microblog owner gave to himself. These data are all downloaded from Sina-Microblog using API. With label '116', through the tags, we can see that people in this community are mostly interested in computer science, and more exactly, they have interests in search engineer, data mining, big data, natural language processing and machine leaning.

4.3 Extract Microblog Tag

We choose the top eight communities in step 4.2, and the number of total user is 205. Using Sina-Microblog API, we download their newly microblog messages including

Table 3. microblog tags of five users

ID	Nickname	Microblog Tag
1792073233	丁国栋_ICT	搜索 市场 架构 重构 搜索引擎 文章
1880650641	张文浩100	网络 数据 挖掘 搜索 社交 算法
1936526225	王斌 ICTIR	数据挖掘 分析 应用 网络 处理 分析
2605098821	范卫卫_大数据	数据 系统 算法 存储 操作系统 搜索
1603279331	张成 ICT	数据 算法 搜索 识别 算法 社交

comments in recent times. The number of messages is 24480, 119 per user. We can get a number of words with ICTCLAS2013 abandoning stopwords. And then get the top tags using the TFIDF ranking algorithm. As is shown in Table 3, there are parts of users' tags we get. The first column is microblogger's id, and the second one is one's nickname. The last one is the tags we get. We can see that the tags of these microbloggers are about computer science, especially data mining and information retrieval. This is corresponding to their research fields.

4.4 Application

As shown in Fig.1, when a query word is submitted, we can get an expansion set. Lucene architecture can help us get the appropriate community items, and then get the proper community. According to the microblog user's influence and the messages time, we can see the related microblog messages in interesting topic. And also, we can get the past microblog messages with the Lucene architecture. All the microblog messages will be stored in database for further research.

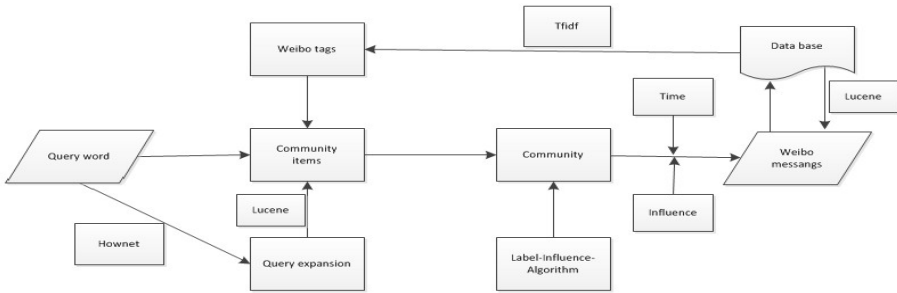


Fig. 1. MSM flow chart

MSM's user interface is inspired by faced search, which has been shown to aid microblog search tasks. As shown in Fig.2. The main screen is a two-column layout. The left column is a list of themes that are related to the current community, while the right column presents the related microbloggers and their microblog tags. As the query word is submitted, the related themes and microblog nicknames appear on the

Microblog Searching Module



Fig. 2. MSM web interface

screen. As themes are selected on the left column, a number of related owners appear at the right column according to their microblog influence. When a microblog nickname is selected, the microblog messages can appear at the right column in chronological order.

5 Conclusion

In this paper, we propose a module named Microblog Searching Module (MSM) to do the microblog search. First, we propose an algorithm named Label-Influence-Algorithm (LIA) to detect communities in microblog network, and then we get the community themes and the microblogger tags using TFIDF ranking algorithm. For a query word, we propose a method to get an expansion set based on the Hownet instead of exact word matching. When a query word is submitted, we can get the related community and then get a list of related microblog owners and themes. But there are also some problems to solve. In the future, firstly, we need to improve the performance of LIA to detect communities in microblog network. Secondly, we have to propose a method to adjust the dynamic network.

Acknowledgement. This work was supported by the Major Project of State Language Commission in the Twelfth Five-year Plan Period (No.ZD1125-1), the NSF of China (No.90920005), the Project in the National Science & Technology Pillar Program in the Twelfth Five-year Plan Period (No.2012BAK24B01), the Program of Introducing Talents of Discipline to Universities (No.B07042), and the NSF of Hubei Province (No.2011CDA034).

References

1. Park, J., Shin, Y., Kim, K., Chung, B.-S.: Searching Social Media Streams on the Web. *IEEE Intelligence Systems* 25(6), 24–31 (2010)
2. Wang, F., Zeng, D., Carley, K., Mao, W.: Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligence Systems* 22(2), 79–83 (2007)
3. Oussalah, M., Bhat, F., Challis, K., Schnier, T.: A Software Architecture for Twitter Collection, Search and Geolocation Services. *Knowledge-Based Systems* 37, 105–120 (2013)
4. Yu, L., Asur, S., Huberman, B.: What Trends in Chinese Social Media. In: *Proceedings of the 5th SNA-KDD Workshop on Web Mining and Social Network Analysis*, San Diego, CA, USA (2011)
5. Anderson, K., Schram, A.: Design and Implementation of a Data Analytics Infrastructure in Support of Crisis Informatics Research. In: *Proceedings of the 33rd International Conference on Software Engineering*, Hawaii, pp. 844–847 (2011)
6. Perera, R., Anand, S., Subbalakshmi, K., Chandramouli, R.: Twitter Analytics: Architecture, Tools and Analysis. In: *Proceedings of Military Communication Conference*, San Jose, CA, pp. 2186–2191 (2010)
7. Connor, B., Krieger, M., Ahn, D.: TweetMotif: Exploratory Search and Topic Summarization for Twitter. In: *Proceedings of the 4th International Conference on Weblogs and Social Media*, Washington, DC, pp. 384–385 (2010)
8. Girvan, M., Newman, M.: Community Structure in Social and Biological Networks. *The National Academy of Science* 99(12), 7821–7826 (2002)
9. Wu, X., Wang, J.: How About Micro-blogging Service in China: Analysis and Mining on Sina Micro-blog. In: *Proceedings of 1st International Symposium on From Digital Footprints to Social and Community Intelligence*, New York, USA, pp. 37–42 (2011)
10. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM* 46(5), 604–632 (1999)
11. Guo, Z., Li, Z., Tu, H., Li, L.: Characterizing User Behavior in Weibo. In: *3rd FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing*, Vancouver, BC, Canada, pp. 60–65 (2012)
12. <http://www.nlp.ir.org/>
13. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24(5), 513–523 (1988)
14. Dong, Z., Dong, Q.: Introduction to Hownet, <http://www.how-net.com>
15. Chen, R., Zhang, L., Lu, C., Mou, L.: Query Expansion Model Based on Concept Graph Information Retrieval. *Journal of Computer Applications* 29(2), 545–548 (2009)

Ambiguity Preferences and Games*

Lan Sun

Maison des Sciences Economiques,
106-112, Boulevard L'Hopital, Paris,
Universite Paris 1-Pantheon Sorbonne
Lan.Sun@malix.univ-paris1.fr

Abstract. Ambiguity (Knightian uncertainty) describes the situations where individuals cannot or do not assign subjective probabilities to uncertain events, in contrast we shall use risk to refer to situations where the decision-maker is familiar with the relevant probabilities. Ellsberg's [1] classic experiments motivate the study of ambiguity. The Ellsberg paradox suggests the standard subjective expected utility theory (SEU) cannot explain such ambiguity preference. Various of preference models are proposed and they are rarely related to one another. This paper interprets the ambiguity aversion from the game theory point of view. We find that the widely used Multiple-Priors reference (MP) can be interpreted by a zero-sum game. We formulate another game to capture a more optimistic ambiguity aversion defined in this paper.

Keywords: Ambiguity, Knightian Uncertainty, Ambiguity Aversion, Zero-sum Game, Preference.

1 Introduction

There are many situations facing to a decision maker that he/she cannot assign a subjective probability distribution to a random variable. For example, from all the information gathered by the investor, he can only estimate that the expected excess return R of an asset belongs to an interval. Then we can assume that the distribution of R belongs to a family of normal distributions $\{N(\mu, \sigma^2) \mid \mu \in [\mu_1, \mu_2]\}$. In contrast, there are some cases, such as the car accidents, health care costs, where the probabilities of the risky events are stable. Knight [2] distinguishes between ambiguity (Knightian uncertainty), where individuals cannot or do not assign subjective probabilities to uncertain events and risk, where the decision-maker is familiar with the relevant probabilities. Ellsberg's classic experiments show that the ambiguity preference is not consistent with the standard Subjective Expected Utility (SEU) theory. Many models of ambiguity preference have been proposed since the Choquet Expected Utility (CEU) approach pioneered by Schmeidler [3]. For complete preferences, besides CEU model, there are Multiple-Priors (MP) model axiomized by Gilboa and Schmeidler [4], the smooth model or KMM-model proposed

* Research supported by Erasmus Mundus Fellowship.

by Klibanoff, et. al [5], as well as multiplier utility and related robust-control-inspired models ([6], [7]). These models are expressed in drastically different formal languages, rarely related to one another.

In this paper, we compare the mainly used complete preference models from the game theory point of view that these preferences differ from the different payoff matrix. We interpreted the utility function in MP model by a zero-sum game where the decision maker has a pessimistic attitude to ambiguity. We define a more optimistic ambiguity preference and formulate a game which can capture this preference. The structure of this paper is following: In section 2, we give a brief review of the Ellsberg’s experiments to conclude the inconsistency of ambiguity aversion with the standard SEU theory. In section 3, we present a zero-sum game to capture the MP model. In section 4, we define an ambiguity preference and formulate the corresponding game to capture it. We explain the games with an example. Conclusion is in section 5.

2 Recall of Ellsberg’s Paradox

Consider the following Ellsberg’s experiments: Suppose there are two urns containing 100 balls each either red or black. Urn A contains 50 red balls and 50 black ones, while there no additional information about urn B. One ball is drawn randomly from each urn. Consider the following acts:

Table 1. Decision under risk

	50	50
	Red	Black
I_A	\$100	\$0
II_A	\$0	\$100

Table 2. Decision under uncertainty

	100	
	Red	Black
I_B	\$100	\$0
II_B	\$0	\$100

Table 3. Decision under risk and uncertainty

	30	60	
	Red	Black	Yellow
I	\$100	\$ 0	\$0
II	\$ 0	\$100	\$0
III	\$100	\$0	\$100
IV	\$0	\$100	\$100

Action I_A is "a bet on red ball drawn from urn A" and II_A is "a bet on black ball drawn from urn A"; Action I_B is "a bet on red ball drawn from urn B" and II_B is "a bet on black ball drawn from urn B" (Table 1, Table 2). Which one do you prefer to bet on, between I_A and II_A , I_B and II_B , I_A and I_B , and II_A and II_B , or are you indifferent? Judging from a large number of responses: I_A and II_A are indifferent, I_B and II_B are indifferent, I_A is preferred to I_B , and II_A is preferred to II_B .

The results from the above experiments indicate that most decision makers are uncertainty averse or ambiguity averse. With these results, we have another Ellsberg experiment yields the violation of the sure-thing principle of Savage postulate. Imagine that there is an urn C known to contain 30 red balls and 60 black and yellow balls. The proportion of yellow balls and black balls are unknown. One ball is to be drawn randomly from the urn. The following actions are considered: Action I is "a bet on red" and action II is "a bet on black". Which action do you prefer? Action III is "a bet on red or yellow" and action IV is "a bet on black or yellow". Which action do you prefer (Table 3)?

A very frequent pattern of response is: action I preferred to II and IV preferred to III . According to Savage's SEU theory, the decision maker will assign a subjective probability to each color ball in the urn. Assume that P_R , P_B , P_Y are the subjective probabilities that the decision maker assigns to red ball, black ball and yellow ball, respectively. If $I \succ II$, then $P_B < P_R$, While $IV \succ III$ means that $P_B + P_Y > P_R + P_Y$, i.e. $P_B > P_R$. These results violate the sure-thing principle of Savage postulate, i. e. if the decision maker prefers action I to II , then he/she should prefer III to IV as well.

3 The Zero-Sum Game Formulation to Multiple-Priors Preference

As we have seen in the previous section that most of the decision makes are ambiguity averse and the ambiguity aversion is not consistent with the standard Subjective Expected Utility (SEU) theory. Many models of ambiguity preference have been proposed, in which the Multiple-Priors (MP) model axiomated by Gilboa and Schmeidler (1989) has been widely used in financial markets. The maxmin expected utility with non-unique priors is:

$$U(X) = \min_{p \in \mathcal{P}} E_p(u(X)) \quad (1)$$

Now let's formulate a Two-player zero-sum game to interpret the MP utility for the Ellsber's experiments, where player 1 chooses the actions and player 2 is the researcher who decides the proportion of the red balls and black balls. Player 2 will pay \$100 to player 1 if player 1 wins, otherwise, both of them will gain nothing and lose nothing. In order to capture the ambiguity aversion property, we assume that the players are risk neutral, and assume that $u(\$100) = 1, u(\$0) = 0$ without loss of generality. G_1 , G_2 , and G_3 are corresponding zero-sum games to the experiments described in Table 1, Table 2, and Table 3, respectively. Player 1 chooses the rows and Player 2 chooses the columns.

Table 4. Zero-sum game G_1 corresponding to decision under risk

		Player 2	
		1/2 Red	1/2 Black
Player 1	I_A	1	0
	II_A	0	1

Table 5. Zero-sum game G_2 corresponding to decision under uncertainty

		Player 2	
		p Red	$1-p$ Black
Player 1	I_B	1	0
	II_B	0	1

Table 6. Zero-sum game G_3 corresponding to decision under risk and uncertainty

		Player 2		
		1/3 Red	p Black	$1-p$ Yellow
Player 1	I	1	0	0
	II	0	1	0
	III	1	0	1
	IV	0	1	1

In game G_1 , the set of moves of Player 1 is $S_A = \{I_A, II_A\}$ and the set of moves of Player 2 is $T_A = \{(1/2, 1/2)\}$. The two strategies of Player 1 $\sigma_1 = (1, 0)$ and $\sigma_2 = (0, 1)$ yield the same payoff which is:

$$U^A(\sigma_1^A) = U^A(\sigma_2^A) = 1/2$$

In game G_2 , $S_B = \{I_B, II_B\}$ and $T_B = \{(p, 1-p) : p \in [0, 1]\}$. Then the payoffs on $\sigma_1^B = (1, 0)$ and $\sigma_2^B = (0, 1)$ of Player 1 are following:

$$U^B(\sigma_1^B) = \min_{\tau} E_{\sigma_1^B, \tau}(g^B(\sigma_1^B, \tau)) = \min_{p \in [0, 1]} p = 0,$$

$$U^B(\sigma_2^B) = \min_{\tau} E_{\sigma_2^B, \tau}(g^B(\sigma_2^B, \tau)) = \min_{p \in [0, 1]} (1-p) = 0,$$

where g is the payoff function corresponding to the payoff matrix. In game G_3 , $S_B = \{I, II, III, IV\}$ and $T_B = \{(1/3, p, 2/3-p) : p \in [0, 1]\}$. Then the payoffs on $\sigma_1^C = (1, 0, 0, 0)$, $\sigma_2^C = (0, 1, 0, 0)$, $\sigma_3^C = (0, 0, 1, 0)$, and $\sigma_4^C = (0, 0, 0, 1)$ of Player 1 are following:

$$U^C(\sigma_1^C) = \min_{\tau} E_{\sigma_1^C, \tau}(g(\sigma_1^C, \tau)) = \min_{p \in [0, 1]} \frac{1}{3} = \frac{1}{3},$$

$$U^C(\sigma_2^C) = \min_{\tau} E_{\sigma_2^C, \tau}(g(\sigma_2^C, \tau)) = \min_{p \in [0, 1]} p = 0,$$

$$U^C(\sigma_3^C) = \min_{\tau} E_{\sigma_3^C, \tau}(g(\sigma_3^C, \tau)) = \min_{p \in [0, 1]} \frac{1}{3} + \frac{2}{3} - p = 0,$$

$$U^C(\sigma_4^C) = \min_{\tau} E_{\sigma_4^C, \tau}(g(\sigma_4^C, \tau)) = \min_{p \in [0, 1]} p + \frac{2}{3} - p = \frac{2}{3}.$$

Comparing these payoffs, we have the results corresponding to the previous three Ellsberg's experiments: σ_1^A and σ_2^A are indifferent, σ_1^B and σ_2^B are indifferent, and σ_1^A is preferred to σ_1^B , σ_1^C is preferred to σ_2^C , but, σ_4^C is preferred to σ_3^C . In order to show that in these games, Player 1 is ambiguity averse in Gilboa and Schmeidler's sense, we need to show that for any two strategies σ_1 and σ_2 are indifferent, then any convex combination of these two strategies are strictly preferred to one of the strategies. Here for $\sigma_1^B \sim \sigma_2^B$, let $\xi \in (0, 1)$,

$$\begin{aligned} U(\xi\sigma_1^B + (1-\xi)\sigma_2^B) &= \min_{\tau} E(g(\xi\sigma_1^B + (1-\xi)\sigma_2^B, \tau)) \\ &= \min_{p \in [0, 1]} (p\xi + (1-p)(1-\xi)) \\ &= \min\{1-\xi, \xi\} \\ &> 0 = \xi U(\sigma_1^B) + (1-\xi)U(\sigma_2^B). \end{aligned}$$

4 The Game Formulation to an Optimistic Ambiguity Aversion

We have justified the zero-sum game can capture the ambiguity aversion behavior defined by MP preference. So the decision makers using MP preference are pessimistic. However, there are many uncertainty scenarios facing to the decision makers are not like a zero-sum game. It will be improper to use MP preference. For example, when an oil company is making decision to drill or not on a site without enough information on the knowledge of the oil capacity in the site. Is the oil company thinking that the nature will play a zero-sum game with him? No, because nature doesn't care whether he wins or loses. In this case, let's assume the payoff matrix as shown in Table 7.

Here the set of strategies of nature (Player 2) is $\{(p, 1 - p) : p \in [0, 1]\}$, where p is the probability of having oil in this site and there is a probability measure μ of p on $[0, 1]$. The set of moves of oil company (Player 1) is $\{Drill, not Drill\}$.

Table 7. Nonzero-sum game G_4 corresponding to decision under uncertainty

		Player 2	
		p	1-p
Player 1	Drill	(2, C)	(-2, C)
	Not Drill	(-1, C)	(0, C)

In this game, the payoff of Player 2 is constant across the states, so all the strategies of Player 2 are indifferent no matter what strategy Player 1 chooses.

Definition. An act X is preferred to Y ($X \succeq Y$), if and only if

$$Pr\{p : \varphi(E_p(u(X))) \geq \varphi(E_p(u(Y)))\} \geq Pr\{p : \varphi(E_p(u(X))) \leq \varphi(E_p(u(Y)))\}, \tag{2}$$

X and Y are indifferent ($X \sim Y$), if and only if

$$Pr\{p : \varphi(E_p(u(X))) \geq \varphi(E_p(u(Y)))\} = Pr\{p : \varphi(E_p(u(X))) \leq \varphi(E_p(u(Y)))\}, \tag{3}$$

X is strictly preferred to Y ($X \succ Y$), if and only if

$$Pr\{p : \varphi(E_p(u(X))) \geq \varphi(E_p(u(Y)))\} > Pr\{p : \varphi(E_p(u(X))) \leq \varphi(E_p(u(Y)))\}, \tag{4}$$

where $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ strictly concave function, and u is a vN-M utility function.

Theorem. If $X \sim Y$, then $\forall \xi \in (0, 1)$, $\xi X + (1 - \xi)Y \succ X \sim Y$, i.e. the decision maker is ambiguity averse with the previous defined preference relations.

Proof.

Assume that $X \succ Y$, $\xi \in (0,1)$,

$$\begin{aligned} & Pr\{p : \varphi(E_p(u(\xi X + (1-\xi)Y))) \geq \varphi(E_p(u(X)))\} \\ & \geq Pr\{p : \varphi(\xi E_p(u(X)) + (1-\xi)E_p(u(Y))) \geq \varphi(E_p(u(X)))\} \\ & > Pr\{p : \xi \varphi(E_p(u(X))) + (1-\xi)\varphi(E_p(u(Y))) \geq \varphi(E_p(u(X)))\} \\ & = Pr\{p : \varphi(E_p(u(Y))) \geq \varphi(E_p(u(X)))\}. \end{aligned}$$

Since u is concave and φ is strictly concave, we can derive the first two inequalities by Jensen's inequality. Similarly, we can have

$$\begin{aligned} & Pr\{p : \varphi(E_p(u(\xi X + (1-\xi)Y))) \leq \varphi(E_p(u(X)))\} \\ & \leq Pr\{p : \varphi(\xi E_p(u(X)) + (1-\xi)E_p(u(Y))) \leq \varphi(E_p(u(X)))\} \\ & < Pr\{p : \xi \varphi(E_p(u(X))) + (1-\xi)\varphi(E_p(u(Y))) \leq \varphi(E_p(u(X)))\} \\ & = Pr\{p : \varphi(E_p(u(Y))) \leq \varphi(E_p(u(X)))\}. \end{aligned}$$

Since X and Y are indifferent, we have

$$\begin{aligned} & Pr\{p : \varphi(E_p(u(Y))) \geq \varphi(E_p(u(X)))\} \\ & = Pr\{p : \varphi(E_p(u(Y))) \leq \varphi(E_p(u(X)))\}. \end{aligned}$$

Finally,

$$\begin{aligned} & Pr\{p : \varphi(E_p(u(\xi X + (1-\xi)Y))) \geq \varphi(E_p(u(X)))\} \\ & \geq Pr\{p : \varphi(E_p(u(\xi X + (1-\xi)Y))) \leq \varphi(E_p(u(X)))\}. \end{aligned}$$

By definition, $\xi X + (1-\xi)Y \succ X \sim Y$.

Q.E.D

How Player 1 will behave in G_4 with this preference? In order to characterize the impact of ambiguity, let's assume that Player 1 is risk neutral and ambiguity averse. Take $u(x) = x$, $\varphi(x) = -\exp\{-\rho x\}$, and μ is a uniform distribution on $[0,1]$. Consider the two strategies of Player 1 $\sigma_1 = (1,0)$ and $\sigma_2 = (0,1)$,

$$\begin{aligned} & Pr\{p : \varphi(E_p(u(g(\sigma_1)))) \geq \varphi(E_p(u(g(\sigma_2))))\} \\ & = Pr\{p : \varphi(4p-2) \geq \varphi(-p)\} = Prob\{p : p \geq \frac{2}{5}\} = \frac{3}{5} \\ & > Pr\{p : \varphi(E_p(u(g(\sigma_1)))) \leq \varphi(E_p(u(g(\sigma_2))))\} = \frac{2}{5}, \end{aligned}$$

where g is the payoff function of G_4 . By definition, Player 1 prefers strategy σ_1 to σ_2 , i.e. he prefers “Drill” to “Not drill” the site. But with MP preference,

$$U^{MP}(\sigma_1) = \min_{\tau} E_{\sigma_1, \tau}(g(\sigma_1, \tau)) = \min_{p \in [0, 1]} 4p - 2 = -2,$$

$$U^{MP}(\sigma_2) = \min_{\tau} E_{\sigma_2, \tau}(g(\sigma_2, \tau)) = \min_{p \in [0, 1]} -p = -1.$$

Player 1 prefers σ_2 to σ_1 , i.e. he prefers “Not drill” to “Drill” the site, which is a pessimistic decision.

5 Conclusion

We have shown that the pessimistic MP preference can be captured by a zero-sum game. The optimal decision to an ambiguity averse decision maker using MP preference is coincide with the optimal strategy of a zero-sum game. In contrast to the pessimistic preference, we formulated a game where the payoffs of the opponent are not influenced by the behavior of the decision maker. The decision maker prefers to the strategy which gives him more chances to get a higher payoff. This preference we defined is more optimistic. A further research on games with ambiguity are encouraged, for example, these preference models only work for solving two outcomes in the Ellsberg’s paradox. Research on preference formulation for three or more outcomes is just in a preliminary stage.

References

1. Ellsberg, D.: Risk, Ambiguity and the Savage Axioms. *Quart. J. Econ.* 75, 643–669 (1961)
2. Knight, F.: *Risk, Uncertainty and Profit*. The Riverside Press, Cambridge (1921)
3. Schmeidler, D.: Subjective Probability and Expected Utility without Additivity. *Econometrica* 57, 571–587 (1989)
4. Gilboa, I., Schmeidler, D.: Maxmin Expected Utility With Non-Unique Prior. *J. Math. Econ.* 18, 141–153 (1989)
5. Klibanoff, P., Marinacci, M., Mukerji, S.: Recursive Smooth Ambiguity Preferences. *J. Econ. Theory* 144, 930–976 (2005)
6. Hansen, L.P., Sargent, T.J.: Recursive Robust Estimation and Control without Commitment. *J. Econ. Theory* 36, 1–27 (2007)
7. Maccheroni, F., Marinacci, M., Rustichini, A.: Dynamic Variational Preferences. *J. Econ. Theory* 18, 141–153 (1989)

Fuzzy Decision Making Based on Fuzzy Propositional Logic with Three Kinds of Negation

Zhenghua Pan

School of Science, Jiangnan University, Wuxi, 214122, China
panzh@jiangnan.edu.cn

Abstract. In [16], negation of fuzzy information were distinguished as contradictory negation, opposite negation and medium negation, and proposed a fuzzy propositional logic with contradictory negation, opposite negation and medium negation (FLCOM). In order to show that FLCOM is applicable for handling fuzzy proposition and its different negations, this paper studied the application of FLCOM to fuzzy decision making in an example. We presented an approach to formal representation of fuzzy proposition and different negations in decision rules, and the measure of truth value of fuzzy proposition and threshold of truth value. Finally, we discussed reasoning and realization on fuzzy decision making in the example based FLCOM and the fuzzy production rule.

Keywords: fuzzy propositional logic, contradictory negation, opposite negation, medium negation, fuzzy decision making.

1 Introduction

The concept of negation plays a special role in information processing, especially in fuzzy information processing. For all of FS (fuzzy sets)[1], RS(rough sets)[2] and the various extensions of FS such as IFS(intuitionistic fuzzy sets)[3], the negation \neg is defined as $\neg x = 1 - x$. In Hájek's BL(basic logic)[4], the negation \neg is defined as $\neg x = x \rightarrow 0$. In Wang's fuzzy propositional calculus system \mathcal{L} [5], the negation \neg is still defined as $\neg x = 1 - x$. Among these theories which exists only a sort of negation, which be simply different forms of expression in definition.

However, the some of scholars have cognized that negation is not a clean concept from a logical point of view in knowledge representation and knowledge reasoning, database query languages (such as *SQL*), production rule systems (such as *CLIPS* and *Jess*), natural language processing and semantic web and so on, there are different negations [6-14]. Wagner et al proposed that there are (at least) two kinds of negation in above domains, a *weak negation* expressing non-truth (in the sense of "she doesn't like snow" or "he doesn't trust you"), and a *strong negation* expressing explicit falsity (in the sense of "she dislikes snow" or "he distrusts you")[6]. Ferré introduced an epistemic extension for the concept of negation, considered that there are *extensional negation* and *intentional negation* in logical concept analysis and natural language [7]. Kaneiwa proposed that "description logic *ALC* with *classical negation* and *strong*

negation, the classical negation \neg represents the negation of a statement, the strong negation \sim may be more suitable for expressing explicit negative information (or negative facts), in other worlds, \sim indicates information that is directly opposite and exclusive to a statement rather than its complementary negation [8]. Since 2005, Pan et al introduced an epistemic extension for the concept of negation, considered that negative relations in knowledge should differentiate the contradictory relation and the opposite relation, and discovered a characteristic of fuzzy information that is if two opposite concepts are fuzzy concepts, there must exists a “medium” fuzzy concept between them, vice versa, thus proposed that negations of fuzzy knowledge include the contradictory negative relation, the opposite negative relation and the medium negative relation, and described these relations using the medium logic[9-12]. In order to provide a set basis and logic basis for contradictory negation, opposite negation and medium negation in fuzzy information, we presented the fuzzy sets with contradictory negation, opposite negation and medium negation (FSCOM)[13][14], and the fuzzy propositional logic with contradictory negation, opposite negation and medium negation (FLCOM), discussed characteristics of FLCOM [16].

In order to show that FLCOM is applicable for handling fuzzy proposition and its different negations, this paper studied the application of FLCOM to fuzzy decision making in an example.

2 Preliminaries

In [16], negation of fuzzy information were distinguished as contradictory negation, opposite negation and medium negation, and proposed a fuzzy propositional logic with contradictory negation, opposite negation and medium negation (FLCOM).

Definition 1. Let S be non-empty set, element in S is called atom proposition or atom formula, the formal symbols “ \neg ”, “ \exists ”, “ \sim ”, “ \rightarrow ”, “ \wedge ” and “ \vee ”, parentheses “(” and “)” are connectives, in which $\neg, \exists, \sim, \rightarrow, \wedge$ and \vee denote contradictory negation, opposite negation, medium negation, implication, conjunction and disjunction. And

(I):

- (a) for any $A \in S$, A is fuzzy formula, formula for short;
- (b) if A and B are formulas then $\neg A, \exists A, \sim A, (A \rightarrow B), (A \vee B)$ and $(A \wedge B)$ are formulas;
- (c) formulas are built from (a) and (b), all the formulas constituted a set $\mathfrak{S}(S)$, \mathfrak{S} for short.

(II): the following formulas in \mathfrak{S} are called axioms:

- (A1) $A \rightarrow (B \rightarrow A)$, (A2) $(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$,
- (A3) $(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$,
- (M₁) $(A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$, (M₂) $(A \rightarrow \exists B) \rightarrow (B \rightarrow \exists A)$
- (H) $\neg A \rightarrow (A \rightarrow B)$, (C) $((A \rightarrow \neg A) \rightarrow B) \rightarrow ((A \rightarrow B) \rightarrow B)$,
- (\vee_1) $A \rightarrow A \vee B$, (\vee_2) $B \rightarrow A \vee B$, (\wedge_1) $A \wedge B \rightarrow A$, (\wedge_2) $A \wedge B \rightarrow B$,
- (Y _{\neg}) $\exists A \rightarrow \neg A \wedge \neg \sim A$, (Y _{\sim}) $\sim A \rightarrow \neg A \wedge \neg \exists A$.

(III): the deduction rule MP(modus ponens): deduce B from $A \rightarrow B$ and A .

(I), (II) and (III) compose a formal system, which is called *fuzzy propositional Logic system with Contradictory negation, Opposite negation and Medium negation*, FLCOM for short.

For the formulas $\neg A$, $\neg\!\!\!\supset A$ and $\sim A$ in FLCOM, the correlation between them can be defined as follows:

Definition 2. In FLCOM,

$$\neg A = \neg\!\!\!\supset A \vee \sim A. \tag{1}$$

In [16], we presented the following semantic interpretation and proved the reliability theorem for FLCOM.

Definition 3 (λ -evaluation). Let $\lambda \in (0, 1)$. Mapping $\partial: \mathfrak{S} \rightarrow [0, 1]$ is called a λ -evaluation if

$$(a) \quad \partial(A) + \partial(\neg\!\!\!\supset A) = 1 \tag{2}$$

$$(b) \quad \partial(\sim A) =$$

$$\left\{ \begin{array}{ll} \lambda - \frac{2\lambda - 1}{1 - \lambda} (\partial(A) - \lambda), & \text{when } \lambda \geq \frac{1}{2} \text{ and } \partial(A) \in (\lambda, 1] \end{array} \right. \tag{3}$$

$$\left\{ \begin{array}{ll} \lambda - \frac{2\lambda - 1}{1 - \lambda} \partial(A), & \text{when } \lambda \geq \frac{1}{2} \text{ and } \partial(A) \in [0, 1 - \lambda) \end{array} \right. \tag{4}$$

$$\left\{ \begin{array}{ll} 1 - \frac{1 - 2\lambda}{\lambda} (\partial(A) + \lambda - 1) - \lambda, & \text{when } \lambda \leq \frac{1}{2} \text{ and } \partial(A) \in (1 - \lambda, 1] \end{array} \right. \tag{5}$$

$$\left\{ \begin{array}{ll} 1 - \frac{1 - 2\lambda}{\lambda} \partial(A) - \lambda, & \text{when } \lambda \leq \frac{1}{2} \text{ and } \partial(A) \in [0, \lambda) \end{array} \right. \tag{6}$$

$$\left\{ \begin{array}{ll} \partial(A), & \text{otherwise} \end{array} \right. \tag{7}$$

$$(c) \quad \partial(A \vee B) = \max(\partial(A), \partial(B)), \quad \partial(A \wedge B) = \min(\partial(A), \partial(B)) \tag{8}$$

$$(d) \quad \partial(A \rightarrow B) = \mathfrak{R}(\partial(A), \partial(B)), \text{ where } \mathfrak{R}: [0, 1]^2 \rightarrow [0, 1] \text{ is a binary function.} \tag{9}$$

3 Application of FLCOM in Fuzzy Decision Making

In order to show that FLCOM is applicable for deal with fuzzy information and its different negations, we take example for fuzzy decision-making in actual life.

Example. In real life, whether an individual (man, family) deposit monthly residuary money in the bank or invest stock, the investment project of the individual lie on the existing savings, the monthly income and Decision Rules. Let x be monthly income of an individual, y be existing savings. Decision Rules:

- a) If y is low savings, then the individual ought to deposit the monthly residuary money in the bank.
- b) If x is high income and y is high savings, then the individual ought to use the monthly residuary money to buy stock.
- c) If x is high income and y is moderate savings, then the individual ought to use the majority of monthly residuary money to invest stock, and deposit the minority of monthly residuary money in the bank.

- d) If x is moderate income and y is moderate savings, then the individual ought to deposit the majority of monthly residuary money in the bank, and use the minority of monthly residuary money to buy stock.

Remark: if x is low income, we think that individual has no residuary money.

Suppose the monthly income of individual M is 5000 (yuan/RMB), the existing savings is 120000. How to establish the investment strategies of M based on decision rules a) – d).

3.1 Formal Representation of Fuzzy Proposition and Its Different Negations in Decision Rules

In real life, people’s viewpoints on the high (or low, moderate) income and the high (or low, moderate) savings are affected by many factors, in which regional difference is main factor. We investigated the people’s viewpoints in some areas of China, obtained the following table (money unit: yuan/RMB):

Table 1. People’s viewpoints on the high (or low) income and the high (or low) savings in some areas of China

<i>views</i>	<i>City/Province</i>	<i>high income</i> (yuan/month)	<i>low income</i> (yuan/month)	<i>high savings</i>	<i>low savings</i>
1	incity/Shanghai	$\geq 15,000$	≤ 2000	$\geq 200,000$	$\leq 100,000$
2	Pudong/Shanghai	$\geq 20,000$	≤ 2500	$\geq 250,000$	$\leq 150,000$
3	Xuhui/Shanghai	$\geq 10,000$	≤ 2000	$\geq 200,000$	$\leq 80,000$
2	Nanjing/Jiangsu	$\geq 10,000$	≤ 1500	$\geq 200,000$	$\leq 80,000$
3	Wuxi/Jiangsu	$\geq 12,000$	≤ 1200	$\geq 150,000$	$\leq 100,000$
4	Suzhou/Jiangsu	$\geq 15,000$	≤ 1500	$\geq 150,000$	$\leq 100,000$
5	Hefei/Anhui	$\geq 6,000$	≤ 1000	$\geq 100,000$	$\leq 80,000$
6	Fuyang/Anhui	$\geq 5,000$	≤ 1000	$\geq 100,000$	$\leq 50,000$
7	Tongning/Anhui	$\geq 4,000$	≤ 800	$\geq 100,000$	$\leq 50,000$
8	Jinan/Shandong	$\geq 7,000$	≤ 1200	$\geq 150,000$	$\leq 80,000$
9	Yantai/Shandong	$\geq 6,000$	≤ 1000	$\geq 120,000$	$\leq 50,000$
10	Weihai/Shandong	$\geq 10,000$	≤ 1500	$\geq 150,000$	$\leq 80,000$

In the table 1, we only listed investigation data in ten cities, and these cities belong to four provinces, respectively. In order to integrate investigation data in the same province, we regard the average of investigation data in the same province as integrative data for each province. Apparently, the more city is investigated for each province, the more nicety for the integrative data. To improve accuracy of integrative data, we take an “elasticity value” for each integrative data respectively, in which the elasticity value is ± 500 /month for “high income”, ± 100 /month for “low income”, ± 20000 for “high savings”, and ± 10000 for “low savings”. We thus got integrative data in each province (Table 2):

Table 2. The integrative data in each province

<i>Province</i>	<i>high income</i> (±500)	<i>low income</i> (±100)	<i>high savings</i> (±20,000)	<i>low savings</i> (±10,000)
Shanghai	≥ 14400	≤ 2000	≥ 210,000	≤ 100,000
Jiangsu	≥ 11000	≤ 1340	≥ 160,000	≤ 82,000
Anhui	≥ 5000	≤ 920	≥ 100,000	≤ 56,000
Shandong	≥ 7000	≤ 1100	≥ 124,000	≤ 68,000

Apparently, in decision rules a) – d)

- “y is low savings”, “y is high savings”, “y is moderate savings”, “x is high income”, “x is moderate income” and “x is low income” are different fuzzy propositions, in which “low savings”, “high savings”, “moderate savings”, “high income”, “low income” and “moderate income” are different fuzzy sets.

According to FLCOM, we need to point out that there is following relationship in these fuzzy propositions:

- the fuzzy proposition “y is low savings” is the opposite negation of “y is high savings”, the fuzzy proposition “y is moderate savings” is the medium negation of the opposite fuzzy propositions “y is low savings” and “y is high savings”.
- the fuzzy proposition “x is low income” is the opposite negation of “x is high income”, the fuzzy proposition “x is moderate income” is the medium negation of the opposite fuzzy propositions “x is low income” and “x is high income”.

Based on FLCOM, we can express different fuzzy proposition in decision rules a) – d) as follows:

- $HIGHincome(x)$, denotes fuzzy proposition “x is high income”
- $\neg HIGHincome(x)$, denotes fuzzy set “x is low income”
- $\sim HIGHincome(x)$, denotes fuzzy set “x is moderate income”
- $HIGHSavings(y)$, denotes fuzzy set “y is high savings”
- $\neg HIGHSavings(y)$, denotes fuzzy set “y is low savings”
- $\sim HIGHSavings(y)$, denotes fuzzy set “y is moderate savings”

where $\neg HIGHincome(x)$ is the opposite negation of $HIGHincome(x)$, $\sim HIGHincome(x)$ is the medium negation of $HIGHincome(x)$ and $\neg HIGHincome(x)$, $\neg HIGHSavings(y)$ is the opposite negation of $HIGHincome(x)$, and $\sim HIGHSavings(y)$ is the medium negation of $HIGHSavings(y)$ and $\neg HIGHSavings(y)$.

In addition, the conclusions in decision rules a) – d) can be express as

$INVESTMENT(stocks)$, denoted individual buy stocks.

$INVESTMENT(savings)$, denoted individual deposit monthly residuary money in the bank.

$MORE(savings, stocks)$, denoted the money which on deposit more than buy stocks.

Thus, the formal representation of decision rules a) - d):

- a) $\neg HIGHSavings(y) \rightarrow INVESTMENT(savings)$
- b) $HIGHincome(x) \wedge HIGHSavings(y) \rightarrow INVESTMENT(stocks)$

- c) $HIGHincome(x) \wedge \sim HIGHsavings(y) \rightarrow INVESTMENT(stocks) \wedge INVESTMENT(savings) \wedge MORE(stocks, savings)$
- d) $\sim HIGHincome(x) \wedge \sim HIGHsavings(y) \rightarrow INVESTMENT(stocks) \wedge INVESTMENT(savings) \wedge MORE(savings, stocks)$

3.2 Measure of Truth Value of Fuzzy Proposition in Decision Rules

In the example, the premises in decision rules a) - d) be composed of aforementioned fuzzy propositions, thus measure of the true value of fuzzy proposition in decision rules is basis of decision making. Based the table 2 and the definition 4, we propose an approach to confirm the truth value of fuzzy proposition as follows.

The table 2 showed that integrative data take on following character: for all of a monthly income datum a , if a belong to the range of high income in Shanghai (that is $a \geq 14400$), then a must belong to the range of high income in other province. If a belong to the range of low income in Anhui (that is $a \leq 920$), then a must belong to the range of low income in other province/city. a be a savings datum as well.

According to this characteristic of integrative data, based on the definition 4 (λ -evaluation ∂), Euclidean distance which $d(x, y) = |x - y|$ and the definition of distance ratio function [15], we can define *truth function* of “ x is high income” and “ y is high savings” as follows.

Definition 7. Let x be any monthly income datum, y be any savings datum. Then, the truth value $\partial(HIGHincome(x))$ of fuzzy proposition “ x is high income” and the truth value $\partial(HIGHsavings(y))$ of fuzzy proposition “ y is high savings” are the following

$$\partial(HIGHincome(x)) = \begin{cases} 0, & \text{when } x \leq \alpha_F + \epsilon_F \\ \frac{d(x, \alpha_F + \epsilon_F)}{d(\alpha_F + \epsilon_F, \alpha_T - \epsilon_T)}, & \text{when } \alpha_F + \epsilon_F < x < \alpha_T - \epsilon_T \\ 1, & \text{when } x \geq \alpha_T - \epsilon_T \end{cases}$$

$$\partial(HIGHsavings(y)) = \begin{cases} 0, & \text{when } y \leq \alpha_F + \epsilon_F \\ \frac{d(y, \alpha_F + \epsilon_F)}{d(\alpha_F + \epsilon_F, \alpha_T - \epsilon_T)}, & \text{when } \alpha_F + \epsilon_F < y < \alpha_T - \epsilon_T \\ 1, & \text{when } y \geq \alpha_T - \epsilon_T \end{cases}$$

where α_T is maximal monthly income datum (or maximal savings datum) for column of *high income* (or *high savings*) in the table 2, ϵ_T is elasticity value. α_F is minimum monthly income datum (or minimum savings datum) for column of *low income* (or *low savings*) in the table 2, and ϵ_F is elasticity value.

By reason that fuzzy proposition “ x is low income” is the medium negation of “ x is high income”, according to (a) in the definition 4, the truth value $\partial(\neg HIGHincome(x))$ of “ x is low income” is the following:

$$\partial(\neg HIGHincome(x)) = 1 - \partial(HIGHincome(x)) \tag{15}$$

Also, by reason that fuzzy proposition “*y is low savings*” is the medium negation of “*y is high savings*”, the truth value $\partial(\neg \text{HIGHsavings}(y))$ of “*y is low savings*” is the following:

$$\partial(\neg \text{HIGHsavings}(y)) = 1 - \partial(\text{HIGHsavings}(y)) \tag{16}$$

According the definition 7, there are the following cases:

(1) If *a* is the monthly income of an individual, since $\alpha_T = 14400$, $\varepsilon_T = 500$, $\alpha_F = 920$ and $\varepsilon_F = 100$, so the truth value $\partial(\text{HIGHincome}(a))$ of fuzzy proposition “*x is high income*” when $x = a$ is the following:

$$\partial(\text{HIGHincome}(a)) = \begin{cases} 0, & \text{when } a \leq 1020 \\ \frac{d(a,1020)}{d(1020,13900)}, & \text{when } 1020 < a < 13900 \\ 1, & \text{when } a \geq 13900 \end{cases} \tag{17}$$

(2) If *b* is existing savings of an individual, since $\alpha_T = 210000$, $\varepsilon_T = 20000$, $\alpha_F = 56000$ and $\varepsilon_F = 10000$, so the truth value $\partial(\text{HIGHsavings}(b))$ of fuzzy proposition “*y is high savings*” when $y = b$ is the following:

$$\partial(\text{HIGHsavings}(b)) = \begin{cases} 0, & \text{when } b \leq 66000 \\ \frac{d(b,66000)}{d(66000,190000)}, & \text{when } 66000 < b < 190000 \\ 1, & \text{when } b \geq 190000 \end{cases} \tag{18}$$

(3) By reason that fuzzy propositions “*x is moderate income*” is the medium negation of “*x is high income*”, and “*y is moderate savings*” is the medium negation of “*y is high savings*”, so we can compute to the truth value $\partial(\sim \text{HIGHincome}(x))$ of “*x is moderate income*” and the truth value $\partial(\sim \text{HIGHsavings}(y))$ of “*y is moderate savings*” according to (b) in the definition 4.

3.3 Threshold of Truth Value Range of Fuzzy Proposition in Decision Rules

Let *A* be a formula in FLCOM. We known that $\lambda(\in(0, 1))$ is an adjustable parameter in the definition 4(λ -evaluation), λ not only reflect the size of range of truth-value $\partial(A)$ of *A* but also reflect the size of ranges of truth-value $\partial(\neg A)$ and $\partial(\sim A)$ of $\neg A$ and $\sim A$, thus show that λ is a “*threshold*” of the truth-value range of a formula in FLCOM. Therefore, the threshold λ is necessary condition to confirm the truth-value range of a formula. We present an approach to confirm threshold of the truth-value range of different fuzzy proposition in decision rules a) - d) as follows.

In the table 2, for all of monthly income datum in Jiangsu, 11000 is minimum (lower limit) for fuzzy set “high income”, 1340 is maximum (upper limit) for fuzzy set “low income”. Hence, for any monthly income datum a_1 ($a_1 \geq 11000$) in Jiangsu, the truth value $\partial(\text{HIGHincome}(a_1))$ of fuzzy proposition “*x is high income*” when $x = a_1$ should take on $\partial(\text{HIGHincome}(a_1)) = 1$. Also, for any monthly income datum a_2 ($a_2 \leq 1340$) in Jiangsu, the truth value $\partial(\neg \text{HIGHincome}(a_2))$ of “*x is low income*” when $x = a_2$ should take on $\partial(\neg \text{HIGHincome}(a_2)) = 1$. However, by (17) and (15), there are

$$\partial(\text{HIGHincome}(11000)) = \frac{d(11000, 1020)}{d(1020, 13900)} = 0.775,$$

$$\partial(\text{HIGHincome}(1340)) = \frac{d(1340, 1020)}{d(1020, 13900)} = 0.025,$$

$$\partial(\neg \text{HIGHincome}(1340)) = 1 - \partial(\text{HIGHincome}(1340)) = 0.975$$

Due to data insufficiency and data distortion, so as to $\partial(\text{HIGHincome}(11000)) \neq 1$ and $\partial(\neg \text{HIGHincome}(1340)) \neq 1$. In order to eliminate this variance, we regard

$$\frac{1}{2}(\partial(\text{HIGHincome}(11000)) + \partial(\text{HIGHincome}(1340)))$$

as threshold λ of the range of truth value $\partial(\text{HIGHincome}(x))$ of “*x is high income*” for any monthly income datum *x* in Jiangsu.

In the same way, we can confirm threshold λ of the range of truth value $\partial(\text{HIGHincome}(x))$ for any monthly income datum *x* in other province in the table 2, and the threshold λ of the range of truth value $\partial(\text{HIGHSavings}(y))$ of fuzzy proposition “*y is high savings*” for any savings datum *y* in each province. Therefore, we got the threshold of the truth value ranges of “*x is high income*” and “*y is high savings*” for each province as follows (Table 3):

Table 3. Threshold λ of the truth value ranges of “*x is high income*” and “*y is high savings*” in each province

Province fuzzy proposition	Jiangsu	Shanghai	Anhui	Shandong
“ <i>x is high income</i> ”	0.875	0.962	0.655	0.729
“ <i>y is high savings</i> ”	0.815	0.863	0.637	0.726

(I) Threshold of truth value range of “*x is high income*” and “*y is high savings*” in decision rules

The above table 3 merely present the threshold of the truth value ranges of fuzzy propositions “*x is high income*” and “*y is high savings*” for four province. In order to get generalization of the threshold, based on the table 3, we regard the average of four values of λ for each fuzzy proposition in the table 3 as respective threshold of the truth value ranges of “*x is high income*” and “*y is high savings*”. Thus, we have the following table (Table 4):

Table 4. Threshold λ of the truth value ranges of “*x is high income*” and “*y is high savings*” in decision rules

fuzzy proposition	“ <i>x is high income</i> ”	“ <i>y is high savings</i> ”
threshold λ	0.805	0.760

- The meaning of threshold λ : for any monthly income datum *a* and savings datum *b*, and the truth value $\partial(\text{HIGHincome}(x))$ of fuzzy proposition “*x is high income*” and the truth value $\partial(\text{HIGHSavings}(y))$ of “*y is high savings*”, if $\partial(\text{HIGHincome}(a)) \geq \lambda$ then *a* belong to the range of High Income, if $\partial(\text{HIGHincome}(a)) \leq 1 - \lambda$ then *a* belong to the

range of Low Income. If $\partial(HIGHsavings(b)) \geq \lambda$ then b belong to the range of High Savings, and if $\partial(HIGHsavings(b)) \leq 1-\lambda$ then b belong to the range of Low Savings. The meaning of threshold λ is showed by the following figure (Fig.1).

(II) *Threshold s of truth value range of “x is low income” and “y is low savings” in decision rules*

According to (15) and (16), the truth value $\partial(\neg HIGHincome(x))$ of “x is low income” and the truth value $\partial(\neg HIGHsavings(y))$ of “y is low savings” are

$$\begin{aligned} \partial(\neg HIGHincome(x)) &= 1-\partial(HIGHincome(x)), \\ \partial(\neg HIGHsavings(y)) &= 1-\partial(HIGHsavings(y)), \end{aligned}$$

by the meaning of threshold in (I), there a belong to the range of Low Income if $\partial(HIGHincome(a)) \leq 1-\lambda$, b belong to the range of Low Savings if $\partial(HIGHsavings(b)) \leq 1-\lambda$, thus, there are a belong to the range of Low Income if $\partial(\neg HIGHincome(a)) \geq \lambda$, and b belong to the range of Low Savings if $\partial(\neg HIGHsavings(b)) \geq \lambda$. Thus, we got threshold s of the truth value ranges of fuzzy propositions “x is low income” and “y is low savings” in decision rules as follows (Table 5):

Table 5. Threshold s of the truth value ranges of “x is low income” and “y is low savings” in decision rules

fuzzy proposition	“x is low income”	“y is low savings”
threshold λ	0.805	0.760

- The meaning of threshold : for any monthly income datum a and savings datum b , and the truth value $\partial(\neg HIGHincome(x))$ of fuzzy propositions “x is low income” and the truth value $\partial(\neg HIGHsavings(y))$ of “y is low savings”, if $\partial(\neg HIGHincome(a)) \geq \lambda$ then a belong to the range of Low Income, if $\partial(\neg HIGHsavings(b)) \geq \lambda$ then b belong to the range of Low Savings. The meaning of threshold λ is showed by the following figure (Fig.1).

(III) *threshold s of truth value range of “x is moderate income” and “y is moderate savings” in decision rules*

The table 4 shown that $\lambda > 1/2$ for threshold λ of the truth value range of fuzzy proposition “x is high income”, $\lambda > 1/2$ for threshold λ of the truth value range of fuzzy proposition “y is high savings”. By the proposition 1, the truth value $\partial(\sim HIGHincome(x))$ of fuzzy proposition “x is moderate income” and the truth value $\partial(\sim HIGHsavings(y))$ of fuzzy proposition “y is moderate savings” satisfy

$$\lambda \geq \partial(\sim HIGHincome(x)) \geq 1-\lambda, \quad \lambda \geq \partial(\sim HIGHsavings(y)) \geq 1-\lambda.$$

Thus, we got the threshold s of the truth value ranges of fuzzy propositions “x is moderate income” and “y is moderate savings” in decision rules as follows (Table 6):

Table 6. Threshold s of the truth value ranges of “x is moderate income” and “y is moderate savings” in decision rules

fuzzy proposition	“x is moderate income”	“y is moderate savings”
threshold $1-\lambda$	0.195	0.240

• The meaning of threshold λ : for any monthly income datum a and savings datum b , and the truth value $\partial(\sim HIGHincome(x))$ of fuzzy proposition “ x is moderate income” and the truth value $\partial(\sim HIGHsavings(y))$ of “ y is moderate savings”, if $\lambda \geq \partial(\sim HIGHincome(a)) \geq 1-\lambda$ then a belong to the range of Moderate Income, if $\lambda \geq \partial(\sim HIGHsavings(b)) \geq 1-\lambda$ then b belong to the range of Moderate Savings. The meaning of threshold λ is showed by the following figure (Fig.1):

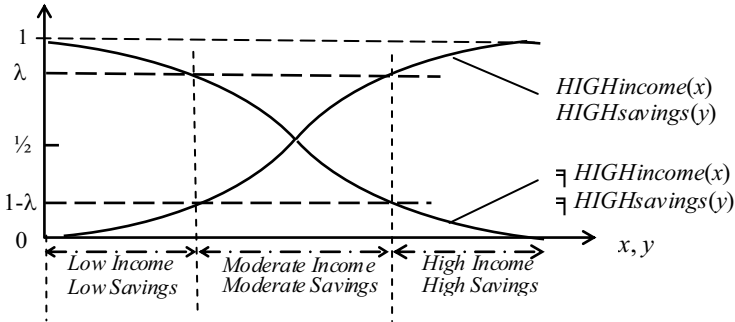


Fig. 1. The meaning of threshold λ of the truth-value range of fuzzy proposition in decision rules

3.4 Reasoning and Decision Making in Example

In decision making rules a)– d), each rule is a fuzzy inference form. How we establish the investment strategy of individual M in the example based on a) – d)? We use fuzzy production rule to discuss investment strategy of individual M. The general form of fuzzy production rule is as follows [14]:

$$P_1, P_2, \dots, P_m \rightarrow Q \mid \langle bd, (\tau_1, \tau_2, \dots, \tau_m) \rangle$$

where P_i ($i = 1, 2, \dots, m$) is fuzzy proposition, which denoted the premise in rule, Q denoted conclusion or actions. bd ($0 \leq bd \leq 1$) is belief degree of rule, τ_i ($0 \leq \tau_i \leq 1, i = 1, 2, \dots, m$) is threshold of the range of truth value $\partial(P_i)$ of P_i . The meaning of fuzzy production rule is that

“deduce Q from P_1, P_2, \dots, P_m with belief degree bd when $\partial(P_i) \geq \tau_i$ for each i ” (19)

Let $bd = 0.9$ (bd can be confirmed by random investigation or field experts). By reason that monthly income of individual M is 5000 yuan, existing savings is 120000 yuan, by (17) and (18), truth value $\partial(HIGHincome(5000))$ of fuzzy proposition “ x is high income” when $x = 5000$ and truth value $\partial(HIGHsavings(120000))$ of fuzzy proposition “ y is high savings” when $y = 120000$ are

$$\begin{aligned} \partial(HIGHincome(5000)) &= \frac{d(5000,1020)}{d(1020,13900)} = 0.309, \\ \partial(HIGHsavings(120000)) &= \frac{d(120000,66000)}{d(66000,190000)} = 0.435. \end{aligned}$$

According to (15) and (16), the truth value $\partial(\neg \text{HIGHincome}(5000))$ of fuzzy proposition “*x is low income*” when $x = 5000$ and truth value $\partial(\neg \text{HIGHSavings}(120000))$ of fuzzy proposition “*y is low savings*” when $y = 120000$ are

$$\begin{aligned}\partial(\neg \text{HIGHincome}(5000)) &= 1 - \partial(\text{HIGHincome}(5000)) = 0.691, \\ \partial(\neg \text{HIGHSavings}(120000)) &= 1 - \partial(\text{HIGHSavings}(120000)) = 0.565.\end{aligned}$$

- For the decision rule a), by reason that premise in a) is fuzzy proposition “*y is low savings*”, according to the table 5, the threshold of truth value range of “*y is low savings*” is 0.760, so a) can be express as

$$\neg \text{HIGHSavings}(y) \rightarrow \text{INVESTMENT}(\text{savings}) \mid \langle 0.9, (0.760) \rangle$$

Due to $\partial(\neg \text{HIGHSavings}(120000)) = 0.565 < 0.760$, by the meaning of threshold in (II), $\partial(\neg \text{HIGHSavings}(120000))$ does not satisfy demand for threshold. Thus, individual M can not adopt the decision rule a) according to (19).

- For the decision rule b), by reason that premise in b) is fuzzy proposition “*x is high income and y is high savings*”, it is a compound fuzzy proposition, which is composed of fuzzy propositions “*x is high income*” and “*y is high savings*”. According to the table 4, the threshold of the truth value range of “*x is high income*” is 0.805, and the threshold of the truth value range of “*y is high savings*” is 0.760, so b) can be express as

$$\text{HIGHincome}(x) \wedge \text{HIGHSavings}(y) \rightarrow \text{INVESTMENT}(\text{stocks}) \mid \langle 0.9, (0.805, 0.760) \rangle.$$

Due to $\partial(\text{HIGHincome}(5000)) = 0.309 < 0.805$ and $\partial(\text{HIGHSavings}(120000)) = 0.435 < 0.760$, by the meaning of threshold in (I), $\partial(\text{HIGHincome}(5000))$ and $\partial(\text{HIGHSavings}(120000))$ does not satisfy demand for threshold. Thus, individual M can not adopt the decision rule b) according to (19).

- For the decision rule c), by reason that premise in c) is fuzzy proposition “*x is high income and y is moderate savings*”, it is a compound fuzzy proposition, which is composed of fuzzy propositions “*x is high income*” and “*y is moderate savings*”. According to the table 4 and table 6, the threshold of the truth value range of “*x is high income*” is 0.805, and the threshold of the truth value range of “*y is moderate savings*” is 0.195, so c) can be express as

$$\begin{aligned}\text{HIGHincome}(x) \wedge \sim \text{HIGHSavings}(y) &\rightarrow \text{INVESTMENT}(\text{stocks}) \wedge \\ \text{INVESTMENT}(\text{savings}) \wedge \text{MORE}(\text{stocks}, \text{savings}) &\mid \langle 0.9, (0.805, 0.195) \rangle.\end{aligned}$$

Due to $\partial(\text{HIGHincome}(5000)) = 0.309 < 0.805$, by the meaning of threshold in (I), $\partial(\text{HIGHincome}(5000))$ does not satisfy demand for threshold. Thus, individual M can not adopt the decision rule c) according to (19).

- For the decision rule d), by reason that premise in d) is fuzzy proposition “*x is moderate income and y is moderate savings*”, it is a compound fuzzy proposition, which is composed of fuzzy propositions “*x is moderate income*” and “*y is moderate savings*”. According to the table 6, the threshold of the truth value range of “*x is high income*” is 0.195, and the threshold of the truth value range of “*y is moderate savings*” is 0.240, so d) can be express as

$$\begin{aligned}\sim \text{HIGHincome}(x) \wedge \sim \text{HIGHSavings}(y) &\rightarrow \text{INVESTMENT}(\text{stocks}) \wedge \\ \text{INVESTMENT}(\text{savings}) \wedge \text{MORE}(\text{savings}, \text{stocks}) &\mid \langle 0.9, (0.195, 0.240) \rangle\end{aligned}$$

For all of the threshold λ of the range of truth value $\partial(\text{HIGHincome}(x))$ of fuzzy proposition “*x is high income*”, both $\lambda = 0.805 > \frac{1}{2}$ in the table 4 and $1-\lambda < \partial(\text{HIGHincome}(5000)) = 0.309 < \lambda$, so there is $\partial(\sim\text{HIGHincome}(5000)) = \partial(\text{HIGHincome}(5000))$ according to (7), that is $\partial(\sim\text{HIGHincome}(5000)) > 0.195$. Thus, $\partial(\sim\text{HIGHincome}(5000))$ satisfied demand for threshold λ . Also, for all of the threshold λ of the range of truth value $\partial(\text{HIGHsavings}(y))$ of fuzzy proposition “*y is high savings*”, both $\lambda = 0.760 > \frac{1}{2}$ in the table 5 and $1-\lambda < \partial(\text{HIGHsavings}(120000)) = 0.435 < \lambda$, so there is $\partial(\sim\text{HIGHsavings}(120000)) = \partial(\text{HIGHsavings}(120000))$ according to (7), that is $\partial(\sim\text{HIGHsavings}(120000)) > 0.195$. Thus, $\partial(\sim\text{HIGHsavings}(120000))$ satisfied demand for threshold λ . By the meaning of threshold λ in (III), $\partial(\sim\text{HIGHincome}(5000))$ and $\partial(\sim\text{HIGHsavings}(120000))$ satisfy demand for each threshold λ . Thus, individual M can adopt the decision rule d) according to (19).

Therefore, the decision rule d) can be as the investment strategy of individual M in the example.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (60973156) and the Fundamental Research Funds for the Central Universities(JUSRP51317B).

References

1. Zadeh, L.A.: Fuzzy Sets. *Information and Control* 8(3), 338–353 (1965)
2. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences* 11, 341–356 (1982)
3. Atanassov, K.: Intuitionistic Fuzzy Sets. *Fuzzy Sets and Systems* 20(1), 87–96 (1986)
4. Hájek, P.: *Metamathematics of Fuzzy Logic*. Kluwer Academic Publishers, Dordrecht (1998)
5. Wang, G.: A fuzzy propositional calculus system. *Bulletin of Science* 42(10), 1041–1045 (1997)
6. Wagner, G.: Web Rules Need Two Kinds of Negation. In: Bry, F., Henze, N., Małuszyński, J. (eds.) *PPSWR 2003. LNCS*, vol. 2901, pp. 33–50. Springer, Heidelberg (2003)
7. Ferré, S.: Negation, Opposition, and Possibility in Logical Concept Analysis. In: Missaoui, R., Schmidt, J. (eds.) *ICFCA 2006. LNCS (LNAI)*, vol. 3874, pp. 130–145. Springer, Heidelberg (2006)
8. Kaneiwa, K.: Description Logic with Contraries, Contradictories, and Subcontraries. *New Generation Computing* 25(4), 443–468 (2007)
9. Pan, Z.: A New Cognition and Processing on Contradictory Knowledge. In: *Fifth International Conference on Machine Learning and Cybernetics*, pp. 1532–1537. Shanghai University Press (2006)
10. Pan, Z., Zhang, S.: Differentiation and Processing on Contradictory Relation and Opposite Relation in Knowledge. In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 334–338. IEEE Computer Society Press (2007)
11. Pan, Z.: Five Kinds of Contradictory Relations and Opposite Relations in Inconsistent Knowledge. In: *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, pp. 761–766. IEEE Computer Society Press (2007)
12. Pan, Z.: A Logic Description on Different Negation Relation in Knowledge. In: Huang, D.-S., Wunsch II, D.C., Levine, D.S., Jo, K.-H. (eds.) *ICIC 2008. LNCS (LNAI)*, vol. 5227, pp. 815–823. Springer, Heidelberg (2008)

13. Pan, Z., Wang, C., Zhang, L.: Three Kinds of Negations in Fuzzy Knowledge and Their Applications to Decision Making in Financial Investment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 391–401. Springer, Heidelberg (2010)
14. Pan, Z., Yang, L., Xu, J.: Fuzzy Set with Three Kinds of Negations and Its Applications in Fuzzy Decision Making. In: Deng, H., Miao, D., Lei, J., Wang, F.L. (eds.) AICI 2011, Part I. LNCS (LNAI), vol. 7002, pp. 533–542. Springer, Heidelberg (2011)
15. Zhang, L., Pan, Z.: Fuzzy Comprehensive Evaluation Based on Measure of Medium Truth Scale. In: 2009 International Conference on Artificial Intelligence and Computational Intelligence, vol. II, pp. 83–87. IEEE Computer Society Press (2009)
16. Pan, Z.: Three Kinds of Negation of Fuzzy Knowledge and their Base of Logic. In: Huang, D.-S., Jo, K.-H., Zhou, Y.-Q., Han, K. (eds.) ICIC 2013. LNCS (LNAI), vol. 7996, pp. 83–93. Springer, Heidelberg (2013)

The Strong Perron Integral of Fuzzy Number Valued Functions

Yabin Shao^{1,2}, Huanhuan Zhang², Yongchun Cao², and Yuqin Bai²

¹ College of Mathematics, Sichuan University, Chengdu 610065, China
yib-shao@163.com

² College of Mathematics and Computer Science Northwest University
for Nationalities Lanzhou 730030, China

Abstract. In this paper, we study the strong Perron integral of fuzzy-number-valued function, and show that the strong fuzzy Perron integral is equivalent to the fuzzy McShane integral.

Keywords: Fuzzy number, strong derivative, strong major function, strong minor function, strong fuzzy Perron integral.

1 Introduction

The major and minor functions are first defined using the upper and lower derivatives, and then the Perron integral is defined using the major and minor functions. The integrals of fuzzy-valued functions have been discussed in recent papers. It is well known, there are three kinds of definitions on fuzzy integrals based on the corresponding backgrounds in integral theory of fuzzy analysis, i.e., Riemann integral, defined by division, sum and limit [4], [5], [6], [7-9], [11]; as an extension of Auman integral for the set-valued functions [1], the integral of fuzzy-valued functions was defined and discussed by Kaleva [6], as an approximation of upper integral and lower integral, the integral was defined by Nanda [8]. However, there exist some drawbacks in Nanda's paper, main reason is it concerned with the supremum and infimum of fuzzy-numbers set. In 1987, Motloka [7] defined a integral used division, sum, upper integral and lower integral, i.e. (M) integral. Nanda [8] characterized Riemann-Stieltjes integral by the same way in 1989. He used the result that 'Bounded fuzzy-number sets have the supremum and infimum, furthermore they could be approached arbitrarily and for continuous functions, they are attainable'. In 1997, Wu Congxin [3] pointed out the mistakes above and proved that the level set of the supremum (infimum) of bounded fuzzy numbers set does not equal to the supremum (infimum) of their level sets. In [14], the Perron integral of fuzzy-valued functions is defined and characterized by using the support functions. In addition, the properties of fuzzy Perron integral are discussed.

In this paper, we change the definitions of major and minor functions by strong derivatives rather than ordinary derivatives, and then define the strong Perron integral of

fuzzy valued functions using such major and minor functions. We also show that the strong Perron integral is equivalent to the McShane integral for fuzzy number valued functions.

2 Preliminary

Let $P_k(R^n)$ denote the family of all nonempty compact convex subset of R^n and define the addition and scalar multiplication in $P_k(R^n)$ as usual. Let A and B be two nonempty bounded subset of R^n . The distance between A and B is defined by the Hausdorff metric [2]:

$$d_H(A, B) = \max \{ \sup_{a \in A} \inf_{b \in B} |a - b|, \sup_{b \in B} \inf_{a \in A} |b - a| \}.$$

Denote $E^n = \{u : R^n \rightarrow [0,1], u \text{ satisfies (1)–(4) below}\}$ is a fuzzy number space. Where

- (1) u is normal, i.e. there exists an $x_0 \in R^n$ such that $u(x_0) = 1$;
- (2) u is fuzzy convex, i.e. $u(\lambda x + (1 - \lambda)y) \geq \min\{u(x), u(y)\}$ for any $x, y \in R^n$ and $0 \leq \lambda \leq 1$,
- (3) u is upper semi-continuous;
- (4) $[u]^0 = cl\{x \in R^n \mid u(x) > 0\}$ is compact.

Define $D : E^n \times E^n \rightarrow (0, +\infty)$

$$D(u, v) = \sup \{ d_H([u]^\alpha, [v]^\alpha) : \alpha \in [0,1] \},$$

where d_H is the Hausdorff metric defined in $P_k(R^n)$. Then it is easy see that D is a metric in E^n . Using the results [3], we know that the metric space (E^n, D) has a linear structure, it can imbedded isomorphically as a cone in a Banach space of function $u^* : I \times S^{n-1}$, where S^{n-1} is the unit sphere in E^n , which an imbedding function $u^* = j(u)$ defined by $u^*(r, x) = \sup_{\alpha \in [u]^\alpha} \langle \alpha, x \rangle$.

Since Hausdorff metric is a kind of stronger metric, much problems could not be characterized. It is well known, the supremum (infimum) is a main concept in analysis, and how to characterized the supremum (infimum) of fuzzy number is an important problem in fuzzy analysis. Refer to (see [2,10]), if $\{u_n\}$ is a bounded fuzzy number sequence, then it has supremum and infimum and if u is supremum (infimum) of $\{u_n\}$, $D(u_n, u) \rightarrow 0$ is not correct generally. The integral metric between two fuzzy numbers by using support functions of fuzzy numbers is defined by Gong (see [12]),

$$D^*(u, v) = \sqrt{\sup_{x \in S^{n-1}, \|x\|=1} \int_0^1 (u^*(r, x) - v^*(r, x))^2 dr}.$$

We easily see that (E^n, D^*) is a metric space, and for each fuzzy number sequence $\{u_n\} \subset E^n$ and fuzzy number $u \in E^n$, if $D(u_n, u) \rightarrow 0$, then $D^*(u_n, u) \rightarrow 0$. The converse result does not hold.

Definition 2.1.[5,6] If a fuzzy number valued function f has H-difference property, i.e., for any $t_1, t_2 \in [a, b]$ satisfying $t_1 < t_2$ there exists a fuzzy number $u \in E^n$, such that $f(t_2) = f(t_1) + u$, u is called H-difference of $f(t_1)$ and $f(t_2)$, we denote $f(t_2) -_H f(t_1) = u$.

Definition 2.2. Let $F : [a, b] \rightarrow E^n$ be a function. The upper and lower Derivates of F at c are defined by

$$\begin{aligned} \overline{D}F(c) &= \limsup_{\delta \rightarrow 0^+} \left\{ \frac{F(x) -_H F(c)}{x - c} : 0 < |x - c| < \delta \right\}; \\ \underline{D}F(c) &= \liminf_{\delta \rightarrow 0^+} \left\{ \frac{F(x) -_H F(c)}{x - c} : 0 < |x - c| < \delta \right\}. \end{aligned}$$

The function F is said to be differentiable at $c \in [a, b]$ if $\overline{D}F(c)$ and $\underline{D}F(c)$ are finite and equal. This common value is called the derivative of F at c and is denoted by $F'(c)$.

Now, we define the strong derivates of a function.

Definition 2.3. Let $F : [a, b] \rightarrow E^n$ be a fuzzy number valued function and let $c \in [a, b]$. The upper and lower strong derivates of F at c are defined by

$$\begin{aligned} \overline{SDF}(c) &= \limsup_{\delta \rightarrow 0^+} \left\{ \frac{F(y) -_H F(x)}{y - x} : [x, y] \subseteq (c - \delta, c + \delta) \cap [a, b] \right\}; \\ \underline{SDF}(c) &= \liminf_{\delta \rightarrow 0^+} \left\{ \frac{F(y) -_H F(x)}{y - x} : [x, y] \subseteq (c - \delta, c + \delta) \cap [a, b] \right\}. \end{aligned}$$

Definition 2.4. A fuzzy number valued function F is said to be strongly differentiable at c if $\overline{SDF}(c)$ and $\underline{SDF}(c)$ are finite and equal. This common value is called the strong derivative of F at c and is denoted by $F'_s(c)$.

Note that the interval $[x, y]$ does not have to contain the point c in the above definition. From definition, it is clear that

$$\underline{SDF} \leq \underline{DF} \leq \overline{DF} \leq \overline{SDF}.$$

From this relation, it is obvious that if F is strongly differentiable at c , then it is differentiable at c and $F'_s(c) = F'(c)$.

The derivative F' of a differentiable function $F : [a, b] \rightarrow E^n$ may not be continuous on $[a, b]$. But the following theorem shows that the strong derivative F'_s of a strongly differentiable function F is in fact continuous on $[a, b]$.

Theorem 2.5. Let $F : [a, b] \rightarrow R$ be a function. If F is strongly differentiable on $[a, b]$, then F'_s is continuous on $[a, b]$.

Proof. Let $c \in [a, b]$ and let $\varepsilon > 0$ be given. Since F is strongly differentiable at c , there exists $\delta > 0$ such that

$$D^* \left(\frac{F(y) -_H F(x)}{y - x}, F'_s(c) \right) < \varepsilon$$

for every interval $[x, y] \subseteq (c - \delta, c + \delta) \cap [a, b]$. If $|z - c| < \delta$ then there exists $\delta_1 > 0$ such that $(z - \delta_1, z + \delta_1) \cap [a, b] \subseteq (c - \delta, c + \delta) \cap [a, b]$ and

$$D^* \left(\frac{F(q) - F(p)}{q - p}, F'_s(z) \right) < \varepsilon \quad \text{for every interval } [p, q] \subseteq (z - \delta_1,$$

$z + \delta_1) \cap [a, b]$, since F is strongly differentiable at z . Choose an interval $[p_0, q_0]$ such that $[p_0, q_0] \subseteq (z - \delta_1, z + \delta_1) \cap [a, b]$. Then we have

$$\begin{aligned} D^* (F'_s(z), F'_s(c)) &\leq D^* \left(F'_s(z), \frac{F(q_0) -_H F(p_0)}{q_0 - p_0} \right) \\ &+ D^* \left(\frac{F(q_0) -_H F(p_0)}{q_0 - p_0}, F'_s(c) \right) < \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

Hence F'_s is continuous at c . This completes the proof.

Definition 2.5. Let $f : [a, b] \rightarrow E^n$ be a fuzzy number valued function, where $R_e = R \cup \{\pm \infty\}$. A measurable function $U : [a, b] \rightarrow E^n$ is called a major function of f on $[a, b]$ if $\underline{DU}(x) > -\infty$ and $\underline{DU}(x) \geq f(x)$ for all $x \in [a, b]$. A measurable function $V : [a, b] \rightarrow R$ is called a minor function of f on $[a, b]$ if $\overline{DV}(x) < \infty$ and $\overline{DV}(x) \leq f(x)$ for all $x \in [a, b]$.

3 The Strong Perron Integral of Fuzzy Valued Functions

Definition 3.1.[14] A fuzzy number valued function $f : [a, b] \rightarrow E^n$ is Perron integrable on $[a, b]$ if f has at least one major function and one minor function on and the fuzzy numbers

$$\inf\{U_a^b : U \text{ is a major function of } f \text{ on } [a, b]\};$$

$$\sup\{V_a^b : V \text{ is a minor function of } f \text{ on } [a, b]\}$$

are equal, where $U_a^b = U(b) -_H U(a)$ and $V_a^b = V(b) -_H V(a)$.

Using upper and lower strong derivatives, we define the strong major and strong minor functions.

Definition 3.2. Let $f : [a, b] \rightarrow E^n$ be a fuzzy number valued function.

(1) A measurable function $U : [a, b] \rightarrow E^n$ is a strong major function of f on $[a, b]$ if $\underline{SDU}(x) > -\infty$ and $\underline{SDU}(x) \geq f(x)$ for all $x \in [a, b]$.

(2) A measurable function $V : [a, b] \rightarrow E^n$ is a strong minor function of f on $[a, b]$ if $\overline{SDV}(x) < \infty$ and $\overline{SDV}(x) \leq f(x)$ for all $x \in [a, b]$.

Now we define the strong fuzzy Perron integral.

Definition 3.3. A fuzzy number valued function $f : [a, b] \rightarrow E^n$ is strongly Perron integrable on $[a, b]$ if f has at least one strong major function and one strong minor function $[a, b]$ and the fuzzy numbers

$$\inf\{U_a^b : U \text{ is a strong major function of } f \text{ on } [a, b]\};$$

$$\sup\{V_a^b : V \text{ is a strong minor function of } f \text{ on } [a, b]\}$$

are equal. This common value is called the strong fuzzy Perron integral of f on $[a, b]$ and is denoted by $(FSP) \int_a^b f d\mu$.

The following theorem is an immediate consequence of the definition.

Theorem 3.4. A fuzzy number valued function $f : [a, b] \rightarrow E^n$ is strongly fuzzy Perron integrable on $[a, b]$ if and only if for each $\varepsilon > 0$ there exist a strong major function U and a strong minor function V of f on $[a, b]$ such that $U_a^b -_H V_a^b < \varepsilon$.

Let $\delta(\cdot)$ be a positive function defined on the interval $[a, b]$. A tagged interval $(x, [a, b])$ consists of an interval $[c, d] \subseteq [a, b]$ and a point $x \in [c, d]$, and a free tagged interval $(x, [c, d])$ consists of an interval $[c, d] \subseteq [a, b]$ and a point $x \in [a, b]$. The (free) tagged interval $(x, [c, d])$ is said to be subordinate to δ if

$$[c, d] \subseteq (x - \delta(x), x + \delta(x)).$$

Let $P = \{(x_i, [c_i, d_i]) : 1 \leq i \leq n\}$ be a finite collection of nonoverlapping tagged intervals in $[a, b]$. If $(x_i, [c_i, d_i])$ is subordinate to δ for each i , then we say that P is subordinate to δ . If P is subordinate to δ and $[a, b] = \bigcup_{i=1}^n [c_i, d_i]$, then we say that P is a (free) tagged partition of $[a, b]$ that is subordinate to δ .

Definition 3.4.[13] A fuzzy number valued function $f : [a, b] \rightarrow E^n$ is McShane integrable on $[a, b]$ if there exists a fuzzy number A with the following property: for every $\varepsilon > 0$ there exists a positive function δ on $[a, b]$ such that $D(f(P) - A) < \varepsilon$, whenever P is a free tagged partition of $[a, b]$ that is subordinate to δ , where $f(P) = \sum_{i=1}^n f(x_i)(d_i - c_i)$ if $P = \{(x_i, [c_i, d_i]) : 1 \leq i \leq n\}$ is a free tagged partition of $[a, b]$. The fuzzy number A is called the fuzzy McShane integral of f on $[a, b]$ and is denoted by $(FM) \int_a^b f d\mu$.

Definition 3.5.[4,11] The fuzzy number valued function $f : [a, b] \rightarrow E^n$ is said to be fuzzy Henstock integrable on $[a, b]$ if we replace ‘free tagged partition’ by ‘tagged partition’ in the definition of the fuzzy McShane integral.

The following two theorems show that the strong fuzzy Perron integral is equivalent to the fuzzy McShane integral.

Theorem 3.6. If $f : [a, b] \rightarrow E^n$ is strongly fuzzy Perron integrable on $[a, b]$, then f is fuzzy McShane integrable on $[a, b]$ and the integrals are equal.

Proof. Let $\varepsilon > 0$ be given. By the definition of strong fuzzy Perron integrability, there exist a strong major function U and a strong minor function V of f on $[a, b]$ such that

$$-\varepsilon < D^* \left(V_a^b, (FSP) \int_a^b f d\mu \right) \leq \tilde{0} \leq D^* \left(U_a^b, (FSP) \int_a^b f d\mu \right) < \varepsilon.$$

Since $\overline{SDV} \leq f \leq \underline{SDU}$ on $[a, b]$, for each $c \in [a, b]$ there exists $\delta(c) > 0$ such that

$$\frac{U(y) -_H U(x)}{y - x} \geq f(c) - \varepsilon \quad \text{and} \quad \frac{V(y) -_H V(x)}{y - x} \leq f(c) + \varepsilon.$$

Now let $P = \{(x_i, [c_i, d_i]) : 1 \leq i \leq n\}$ be a free tagged partition of $[a, b]$ that is subordinate to δ . Then we have

$$\begin{aligned} & D^* \left(\sum_{i=1}^n f(x_i)(d_i - c_i), (FSP) \int_a^b f \, d\mu \right) \\ &= D^* \left(\sum_{i=1}^n f(x_i)(d_i - c_i), U_a^b \right) + D^* \left(U_a^b, (FSP) \int_a^b f \, d\mu \right) \\ &< \sum_{i=1}^n \varepsilon(d_i - c_i) + \varepsilon = \varepsilon(b - a + 1). \end{aligned}$$

Similarly, using the strong minor function V ,

$$D^* \left(\sum_{i=1}^n f(x_i)(d_i - c_i), (FSP) \int_a^b f \, d\mu \right) > -\varepsilon(b - a + 1).$$

Since $D^* \left(f(P), (FSP) \int_a^b f \, d\mu \right) < \varepsilon(b - a + 1)$, f is fuzzy McShane integrable on $[a, b]$ and $(FM) \int_a^b f \, d\mu = (FSP) \int_a^b f \, d\mu$.

Definition 3.6.[4,11] Let $\omega(F, [c, d]) = \sup\{|F(y) - F(x)|\}$ denote the oscillation of the fuzzy number valued function F on the interval $[c, d]$. The function F is BV_* on E if $\sup\{\sum_{i=1}^n \omega(F, [c_i, d_i])\}$ is finite, where the supremum is over all finite collections $\{[c_i, d_i] : 1 \leq i \leq n\}$ of non-overlapping intervals in $[a, b]$. And the function F is said to be BVG_* on E if E can be written as a countable union of sets on each of which F is BV_* .

Theorem 3.7. If a fuzzy number valued function $f : [a, b] \rightarrow E^n$ is fuzzy McShane integrable on $[a, b]$, then f is strongly fuzzy Perron integrable on $[a, b]$.

Proof. Let $\varepsilon > 0$ be given. By the definition of McShane integrability, there exists a positive function δ on $[a, b]$ such that $D^* \left(f(P), (FM) \int_a^b f \, d\mu \right) < \varepsilon$

whenever P is a free tagged partition of $[a, b]$ that is subordinated to δ . For each $x \in (a, b]$, let

$$U(x) = \sup \{ f(P) : P \text{ is a free tagged partition of } [a, x] \};$$

$$V(x) = \inf \{ f(P) : P \text{ is a free tagged partition of } [a, x] \};$$

and let $U(a) = 0 = V(a)$. Then we have the functions U and V are finite-valued on $[a, b]$. We prove that U is a strong major function of f on $[a, b]$; the proof that V is a strong minor function of f on $[a, b]$ is quite similar.

Fix a point $c \in [a, b]$ and let $[x, y]$ be any interval such that $[x, y] \subseteq (c - \delta(c), c + \delta(c)) \cap [a, b]$. For each free tagged partition P of $[a, x]$ that is subordinate to δ , we find that $U(y) \geq f(P) + f(c)(y - x)$ and it follows that $U(y) \geq U(x) + f(c)(y - x)$. This shows that $\frac{U(y) -_H U(x)}{y - x} \geq f(c)$

and hence $\underline{SDU}(c) \geq f(c) > -\infty$. Since $-\infty < \underline{SDU} \leq \underline{DU}$ on $[a, b]$, U is BVG_* on $[a, b]$ and it follows that U is measurable on $[a, b]$. Hence, U is a strong major function of f .

Since $D^*(f(P_1), f(P_2)) < 2\epsilon$ for any two free tagged partitions P_1 and P_2 of $[a, b]$ that are subordinate to δ , it follows that $U_a^b -_H V_a^b \leq 2\epsilon$. By Theorem 3.4, the function f is strongly fuzzy Perron integrable on $[a, b]$

Acknowledgments. This work is partially supported by National Natural Science Foundation of China under grant No. 11161041, and 2012 Scientific Research Fund for State Ethnic Affairs Commission of China, and Fundamental Research Fund for the Central Universities No. 31920130005, No. 31920130009 and No. 31920130010.

References

1. Aumann, R.: Integrals of Set-Valued Functions. *J. Math. Anal. Appl.* 12, 1–12 (1965)
2. Zhang, B.K.: On Measurability of Fuzzy-Number-Valued Functions. *Fuzzy Sets and Syst.* 120, 505–509 (2001)
3. Wu, C.X., Wu, C.: The Supremum and Infimum of the Set of Fuzzy Numbers and Its Application. *J. Math. Anal. Appl.* 210, 499–511 (1997)
4. Wu, C.X., Gong, Z.T.: On Henstock Integral of Fuzzy-Number-Valued Functions (I). *Fuzzy Sets and Syst.* 120, 523–532 (2001)
5. Goetschel, R., Voxman, W.: Elementary Fuzzy Calculus. *Fuzzy Sets and Syst.* 18, 31–43 (1986)

6. Kaleva, O.: Fuzzy Differential Equations. *Fuzzy Sets and Syst.* 24, 301–317 (1987)
7. Matloka, M.: On Fuzzy Integral. In: *Proc. 2nd Polish Symp. on Interval and Fuzzy Math.*, pp. 167–170. Wydawnicacwo Politechniki Poznanskiej, Poznan (1987)
8. Nanda, S.: On fuzzy integrals. *Fuzzy Sets and Syst.* 32, 95–101 (1989)
9. Wu, H.C.: Fuzzy-Valued Integral of Fuzzy-Valued Measurable Functions With Respect to Fuzzy-Valued Measures Based on Closed Intervals. *Fuzzy Sets and Syst.* 87, 65–78 (1997)
10. Xue, X.P., Ha, M.H.: Random Fuzzy Number Integrals in Banach Spaces. *Fuzzy Sets and Syst.* 66, 97–111 (1994)
11. Gong, Z.T.: On Henstock Integral of Fuzzy-Number-Valued Functions (II): the Descriptive Characteristics of (FH) Integral and Its Convergence Theorems. *Fuzzy Sets and Syst.* 145, 381–393 (2004)
12. Gong, Z.T., Wu, C., Cheng, D.G.: The Partial Ordering, Supremum and Infimum and Metric Description of the Space of Fuzzy Numbers. *Journal of Lanzhou University* (to appear)
13. Gong, Z.T., Wu, C.X.: The McShane Integral of Fuzzy-Valued Functions. *Southeast Asian Bull. of Math.* 24, 365–373 (2000)
14. Gong, Z.T., Wang, L., Shao, Y.B.: The Perron Integral of Fuzzy Valued Functions: Background, Definition, and Characterization. *Int. J. of Pure and Applied Math.* 3, 321–336 (2005)

Similarity of Fuzzy Triangular Number Based on Indifference Area and Its Application

Xixiang Zhang¹ and Jianxun Liu²

¹ Business School, Jiaxing University, Jiaxing, China

² Library, Jiaxing University, Jiaxing, China
zhangmiddle@126.com

Abstract. Fuzzy triangular number was applied in area such as risk assessment and performance evaluation because it can better express fuzzy linguistic terms. There were several approaches to measure the similarity of fuzzy triangular numbers. A new method to measure fuzzy triangular number was put forward based on the location and area of two fuzzy triangular numbers. And it was applied to collaborative filtering recommendation system to measure users' similarity. A case was used to illustrate the proposed method and the result showed that it can obtain better discriminability.

Keywords: triangular fuzzy number, similarity, discriminability.

1 Introduction

Fuzzy triangular number can better express fuzzy linguistic terms, it has been integrated with multi-attribute decision-making [1], which has been applied in fields such as risk evaluation and performance evaluation.

Similarity is used to measure the similar degree of two fuzzy numbers. Chen and Lin proposed the distance of two fuzzy triangular numbers and calculated the similarity of two fuzzy triangular based on the distance [2]. Hsieh and Chen put forward the concept of graded mean integration representation and computed the similarity of two fuzzy triangular numbers based on it [3]. Yager used the kernel function to represent a fuzzy triangular number [4], and Xu defined the concept of fuzzy expectation value of a fuzzy number [1].

The overlap part of two fuzzy triangular numbers is indifferent, and the ratio of the overlap part can be used to represent the similarity of two fuzzy triangular numbers. And the fuzzy triangular number can express the user's comprehensive evaluation of an item in collaborative filter recommendation system. The similarity represented by the ratio of the overlap part can express the similarity of two users using the fuzzy triangular numbers. The comparison of users' similarity was carried out and the result showed that the fuzzy triangular number could express the user's comprehensive evaluation on items, it can obtain better discriminability.

2 Fuzzy Triangular Numbers

If a regular convex fuzzy set \tilde{M} based on the real domain and meets the following conditions: (1) there exists only one element x_0 that $\mu_{\tilde{M}}(x_0) = 1$; (2) $\mu_{\tilde{M}}(x)$ is continuous. Then \tilde{M} is a fuzzy number, which means “a real number that be approximate to x_0 ”. \tilde{M} can be expressed as follows[5].

$$\mu_{\tilde{M}}(x) = \begin{cases} L(x), l \leq x \leq m \\ R(x), m \leq x \leq r \end{cases} \tag{1}$$

$L(x)$ is a increasing function and continuous on the right, $R(x)$ is a decreasing function and continuous on the left and $0 \leq L(x), R(x) \leq 1$. When $L(x)$ is expressed as $\frac{x-l}{m-l}$ and $R(x)$ is represented as $\frac{x-r}{m-r}$, $\mu_{\tilde{M}}(x)$ is called a fuzzy triangular number. To donate it simply, the left threshold value a^L , the midpoint a^M and the right threshold value a^U are used to represent a fuzzy triangular number \tilde{A} , $\tilde{A} = (a^L, a^M, a^U)$, and its membership function is as follows.

$$\mu_{\tilde{M}}(x) = \begin{cases} \frac{x - a^L}{a^M - a^L}, a^L \leq x \leq a^M \\ \frac{x - a^U}{a^M - a^U}, a^M \leq x \leq a^U \\ 0, otherwise \end{cases} \tag{2}$$

3 Similarity of Fuzzy Triangular Numbers Based on Indifference Area

The more the indifference part of two fuzzy triangular numbers, the more similar between the shape and location of them, and the bigger the similarity of them should be. The similarity of two fuzzy triangular numbers can be expressed by the ratio of the indifference part of them. Let $\tilde{A} = (a^L, a^M, a^U)$ and $\tilde{B} = (b^L, b^M, b^U)$ are two fuzzy triangular fuzzy numbers, the similarity of them can be donated as follows.

$$SC_{\tilde{A}\tilde{B}} = \frac{2A_{\tilde{A}\cap\tilde{B}}}{A_{\tilde{A}} + A_{\tilde{B}}} \tag{3}$$

$A_{\tilde{A} \cap \tilde{B}}$ is the area of the indifference part of two fuzzy triangular numbers, $A_{\tilde{A}}$ and $A_{\tilde{B}}$ is the area of \tilde{A} and \tilde{B} respectively, and $A_{\tilde{A}} = \frac{1}{2}(a^U - a^L)$.

The area of the indifferent part of two fuzzy triangular numbers is calculated based on the cross points of the fuzzy triangular numbers. Each fuzzy triangular number can be expressed as a left increasing line and a right decreasing line on the coordinate axis. The lines of fuzzy triangular number \tilde{A} and \tilde{B} can be donated as follows.

$$\begin{aligned}
 l_{\tilde{A}} &: x = (a^M - a^L)y + a^L \\
 r_{\tilde{A}} &: x = (a^M - a^U)y + a^U \\
 l_{\tilde{B}} &: x = (b^M - b^L)y + b^L \\
 r_{\tilde{B}} &: x = (b^M - b^U)y + b^U
 \end{aligned}
 \tag{4}$$

$l_{\tilde{A}}$ can intersect with $l_{\tilde{B}}$ and $r_{\tilde{B}}$, $r_{\tilde{A}}$ can be intersect with $l_{\tilde{B}}$ and $r_{\tilde{B}}$. The number of intersecting points of these four lines can not exceed 3. So the intersecting situation of two fuzzy triangular numbers can be concluded as figure1.

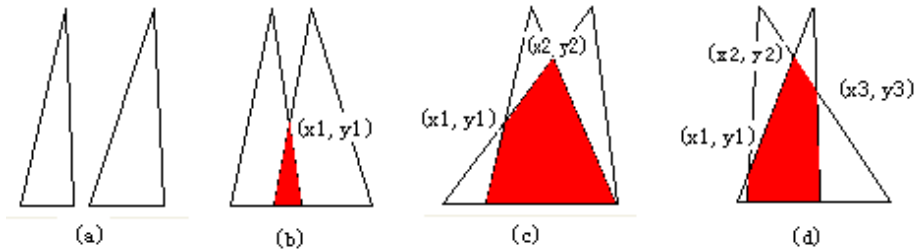


Fig. 1. the intersecting situation of two fuzzy triangular numbers

Obviously, the area of the indifferent part is 0 if there is no crossing point. If there is one crossing point (x_1, y_1) , the area of the indifferent part is obtained from the following formula.

$$A_{\tilde{A} \cap \tilde{B}} = 0.5(a^U - b^L) * y_1
 \tag{5}$$

If there are two crossing points which are donated as (x_1, y_1) and (x_2, y_2) , the area of the indifferent part is follows.

$$\begin{aligned}
 A_{\tilde{A} \cap \tilde{B}} &= 0.5((x_1 - \max(a^L, b^L)) * y_1 + \\
 &(y_1 + y_2)(x_2 - x_1) + (\min(a^U, b^U) - x_2) * y_2)
 \end{aligned}
 \tag{6}$$

If there are three crossing points that are donated as (x_1,y_1) , (x_2,y_2) and (x_3,y_3) , the area of the indifferent part is obtained from the following formula.

$$A_{\tilde{A} \cap \tilde{B}} = 0.5((x_1 - \max(a^L, b^L)) * y_1 + (y_1 + y_2)(x_2 - x_1) + (y_3 + y_2)(x_3 - x_2) + (\min(a^U, b^U) - x_3) * y_3) \tag{7}$$

The above method to measure the similarity of two fuzzy triangular numbers does not take the distance of two fuzzy triangular numbers into account. When two fuzzy triangular numbers are close to each other and have small uncertain range, the similarity may be 0. Therefore, expectation value is used.

$$S_{\tilde{A}, \tilde{B}} = \alpha SC_{\tilde{A}, \tilde{B}} + \beta SE_{\tilde{A}, \tilde{B}}, \alpha + \beta = 1 \tag{8}$$

The similarity based on expectation value is obtained from the following formula.

$$SE_{\tilde{A}, \tilde{B}} = 1 - D(\tilde{A}, \tilde{B}) \tag{9}$$

$$D(\tilde{A}, \tilde{B}) = | (a^L + 2a^M + a^U) / 4 - (b^L + 2b^M + b^U) / 4 |$$

4 Application in Collaborative Filter Recommendation

Collaborative filtering recommendation system based on nearest neighbors is the most successful recommendation technology. Its basic principle is: the users will have similar evaluation on other items if users have similar evaluation on some items. Therefore, similarity measurement based on users' evaluation on items is the key problem in collaborative filtering recommendation systems. comprehensive measurements such as evaluation ranking frequency vector, could model are used to represent users' comprehensive evaluation and to calculate users' similarity[6-8].

Take the example from Zhang[8], there are four users, which are donated as A,B,C and D, to evaluate 10 items. Table 1 shows the detailed evaluation.

Table 1. Users' evaluation on 10 items

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
User_A	2	1	1	1	2	1	1	2	1	2
User_B	5	4	5	4	5	4	5	4	5	4
User_C	4	5	3	4	5	5	4	4	5	3
User_D	2	1	2	2	1	1	2	2	1	2

Zhang[8] found the similarity based on evaluation ranking frequency vector was in [0.98,1] and it's difficult to distinguish its nearest neighbors. They used the cloud model to express the user's evaluation on items, and got the similarity of users based on cloud model. Table 2 shows the result.

Table 2. Users' similarity based on cloud model

	A	B	C	D
A	1	0.956	0.965	0.999
B	0.956	1	0.999	0.967
C	0.965	0.999	1	0.975
D	0.999	0.967	0.975	1

The fuzzy triangular number show the fuzzy and uncertainty of information, it is appropriate to represent the user's comprehensive evaluation on items. The average of the user's comprehensive evaluation on items represents the midpoint of a fuzzy triangular number, which means the degree of concentration on items. The absolute value of user's evaluation on item minus the average of user's evaluation on items is called the fuzzy discretization (FD) of user's evaluation, which means the uncertainty of user's evaluation. Therefore, if a user evaluates the items in $X = \{x_1, x_2, \dots, x_n\}$, its comprehensive evaluation can be expressed by a fuzzy

triangular number $\tilde{A} = (a^L, a^M, a^U)$.

$$\begin{aligned}
 a^M &= \frac{1}{5n} \sum_{i=1}^n x_i \\
 a^L &= (a^M - FD) / 5 \\
 a^U &= (a^M + FD) / 5 \\
 FD &= \frac{1}{5n} \sum_{i=1}^n |x_i - a^M|
 \end{aligned} \tag{10}$$

According to the above formula, the four users' comprehensive evaluation on items can be represented by the following fuzzy triangular numbers respectively: $\tilde{A} = (0.184, 0.28, 0.376)$, $\tilde{B} = (0.8, 0.9, 1)$, $\tilde{C} = (0.712, 0.84, 0.928)$ and $\tilde{D} = (0.224, 0.32, 0.416)$. The similarity of four users is obtained, table 3 shows the result.

Table 3. Users' similarity based on fuzzy triangular numbers

	A	B	C	D
A	1	0.304	0.363	0.9
B	0.304	1	0.826	0.336
C	0.363	0.826	1	0.395
D	0.9	0.336	0.395	1

User A and B have different evaluation on the ten items, the similarity based on cloud model is 0.956, which is not consistent with our intuitionistic judging. While the similarity based on fuzzy triangular numbers is 0.304, which means that there is big difference between the evaluation of user A and B. Therefore, it can be concluded that it is reasonable to use a fuzzy triangular number to express a user's evaluation on items and calculate the users' similarity based on it.

5 Conclusion

Fuzzy triangular number can express the vagueness and uncertainty of information, it is easy to understand. It has been applied to many fields such as risk analysis and human resource evaluation. The similarity of two fuzzy triangular numbers based on expectation value is always large. To overcome this shortcoming, a new method to measure the similarity of two fuzzy triangular numbers is proposed, which is measured based on the indifferent area.

The similarity of fuzzy triangular numbers was applied to collaborative filtering recommendation system. A case was used to calculate the users' similarity based on cloud model and fuzzy triangular numbers. The conclusion shows that the fuzzy triangular numbers can be used to represent the users' evaluation on items and get better discriminability.

Acknowledgment. The paper was supported by The Ministry of Education of Humanities and Social Sciences Youth Fund(Granted by 12YJCZH284)

References

1. Xu, Z.S.: Method Based on Expected Values for Fuzzy Multiple Attribute Decision Making Problems with Preferences Information on Alternatives. *Syst. Eng. Theory & Practice* 1, 109–113 (2004)
2. Chen, S.M., Lin, S.Y.: A New Method for Fuzzy Different Generalized Fuzzy Numbers. In: *Proc. 1995 Artificial Intelligence Workshop*, pp. 245–250 (1995)
3. Hsieh, C.H., Chen, S.H.: Similarity of Generalized Fuzzy Numbers with Graded Mean Integration Representation. In: *Proc. 8th Int. Fuzzy Systems Association World Cong.*, pp. 551–555
4. Yager, R.R.: A Procedure for Ordering Fuzzy Subsets of the Unit Interval. *Inform. Sciences* 1, 143–161 (1981)
5. Cheng, L.C., Chen, H.C.: Operators of Fuzzy Triangular Numbers Based on Bounded Operator. *Fuzzy Syst.* 2, 33–42 (1993)
6. Choi, K.H., Suh, Y.M.: A New Similarity Function for Selecting Neighbors for Each Target Item in Collaborative Filtering. *Knowl.-Based Syst.* 37, 146–153 (2013)
7. Bellogín, A., Cantador, I., Castells, P.: A Comparative Study of Heterogeneous Item Recommendations in Social Systems. *Inform. Sciences* 221, 142–169 (2013)
8. Zhang, G.W., Li, D.Y., Li, P., Kang, J.C., Chen, G.S.: A Collaborative Filtering Recommendation Algorithm Based on Cloud Model. *J. Softw.* 18(10), 2403–2411 (2007)

Research on Image Mosaic Algorithm Based on Computer Wizard Vector Field Algorithm

Xiaobo Gao and Xianmei Fang

Department of Computer and Information Science Hechi University
{aspone, 99465494}@cq.com

Abstract. With the development of computer science and technology, image mosaic technology has been widely used in various fields including satellite remote sensing technology, image magnification and synthesis technology. Image magnification and synthesis is one of the most commonly used technologies in image Mosaic technology. In the image magnification and synthetic process, it often appears smaller resolution and distortion of image registration part. This paper uses computer vector field theory to establish the image Mosaic algorithm, structures the vector control model of pattern synthesis and amplification, simulates image Mosaic algorithm through the MATLAB simulation module ,gets the residual graph of image synthesis and effect of image magnification and synthesis, finds that computer vector algorithm can increase the resolution of the image after magnification through the effect comparison and synthetic image contact ratio is better which proves the effectiveness of computer vector field algorithm in image synthesis technology.

Keywords: computer, wizard vector field, image Mosaic, simulation, vector control, contact ratio.

1 Introduction

The application field of image mosaic technology is very wide. Image Mosaic algorithm is the foundation to realize image Mosaic technology. It is not easy to get algorithm of high efficiency and stability. This paper develops image Mosaic algorithm of mathematical model based on the vector field algorithm, establishes image artwork vector matrix, coordinate transformation matrix and coordinate matrix after image Mosaic [1,2]. It finally gets the amplification results after general amplification and vector field calculation through the image simulation using MATLAB simulation module. It finds the validity and feasibility of the image Mosaic algorithm through the analysis of the results. Finally the article puts forward the steps of wizard vector field algorithm which provides a theoretical reference for the development of image Mosaic technology.

2 The Wizard Vector Field Algorithm of Image Mosaic Technology

Image Mosaic technology is achieved mainly through the vector transformation of the image coordinates. So the vector coordinate transformation is the main research object of image Mosaic algorithm [3-6]. Image Mosaic algorithm guide vector field can be divided into three coordinate systems, one is the original vector coordinate system, and the other is the vector algorithm coordinate system and the Mosaic image vector coordinate system. Original vector coordinate system transforms the coordinates through the vector wizard coordinate system and finally obtains the image Mosaic vector coordinate system and realizes image Mosaic process. The specific process is shown in Figure 1.

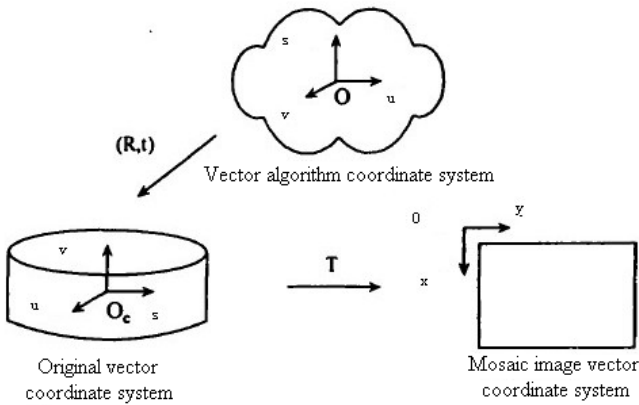


Fig. 1. The process of image coordinates vector transformation

The image synthesis directional control can be realized through the image coordinate resolution model operation. The basic equation of operation is as follows [7].

$$\begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix} = \begin{bmatrix} K_1 & 0 & K_2 & 0 \\ 0 & K_1 & 0 & K_2 \\ K_2 S & Z_1 K_2 & S & Z_1 K_3 \\ 0 & K_2 & 0 & K_3 \end{bmatrix} \begin{pmatrix} iV_1 \\ iV_2 \\ iV_3 \\ iV_4 \end{pmatrix} \quad (1)$$

In this equation, V_1, V_2, V_3, V_4 are resolution mode positions and time coordinates. K_1 is coordinate after lengthen. K_2 is coordinate after widen. K_3 is coordinate after heighten. Z_1 is resolution base resolution. Z_2 is resolution Angle resolution. S is differential operator. U_1, U_2 are resolution profile control.

Image synthesis equation is:

$$\begin{pmatrix} H_1 \\ H_2 \\ H_3 \\ H_4 \end{pmatrix} = \begin{bmatrix} U_1 + K_1S & 0 & K_2S & 0 \\ 0 & U_1 + K_1S & 0 & K_2S \\ K_2S & Z_1K_2 & R_f + K_3S & Z_1K_3 \\ -Z_1K_2 & -K_2 & 0 & U_2 + K_3S \end{bmatrix} \begin{pmatrix} iV_1 \\ iV_2 \\ iV_3 \\ iV_4 \end{pmatrix} \quad (2)$$

In formula (2), H_1, H_2, H_3, H_4 are Image synthesis vectors. Firstly, we can transfer them to the K - S coordinate of stillness through the $3/2$ transformation then get resolution and Mosaic image component through the rotating space vector coordinates. This will realize decoupling control of image and speed up the image Mosaic. Image perspective projection equation can be shown as follows [8].

$$\frac{A'}{A} = \frac{B'}{B} = \frac{S}{C} \quad (3)$$

In 3D field, the plane coordinates of spot can be shown as:

$$A' = \frac{S}{C}A \quad B' = \frac{S}{C}B \quad (4)$$

The transformation relationship of image Mosaic wizard vector field coordinate is [9]:

$$C \begin{bmatrix} A' \\ B' \\ 1 \end{bmatrix} = \begin{bmatrix} S & 0 & 00 \\ 0 & S & 00 \\ 0 & 0 & 10 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ 1 \end{bmatrix} \quad (5)$$

3 Simulation Calculation of MATLAB Image Mosaic

In order to validate the effectiveness of the algorithm in the first part of the vector field, we need to use numerical simulation software. MATLAB is professional simulation software. Its simulation module can realize perfect simulation effect. This paper uses MATLAB software to simulate vector field image Mosaic algorithm. It gets the data by vector transformation and establishes the coordinate values of simulation image and input the data into MATLAB procedure. Vector table is shown in Table 1.

Table 1. Vector coordinate table

Image category	X	Y	Z	T(Time axis)
Original drawing	(5,8,19...)	(3,5,8...)	(2,4,7,...)	(1,2,3...)
Conversion coordinates	(1,5,3...)	(2,6,9...)	(3,6,8...)	(1,2,3...)

In image synthesis process, we need to synthesize variety colors of image. We can get the MATLAB image synthesis simulation color synthetic residual effect through simulation calculation which is shown in Figure 2.

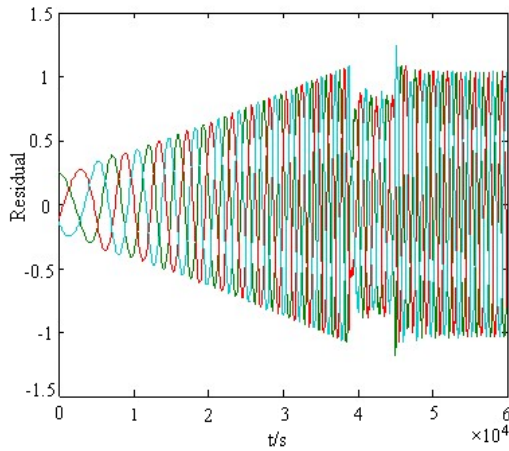


Fig. 2. The process of image coordinates vector transformation image synthesis convergence curve

We can see from Figure 2 that in the start stage of convergence curve, contact ratio of three color curves is poorer; image synthesis effect is not good. The image synthesis reclosing part residual is smaller and smaller after computer vector field iterative algorithm and finally reach the ideal image Mosaic synthesis effect.



Fig. 3. Definition profile in image magnification process

From Figure 3, we can see that the image ordinary amplification process will appeared image resolution distortion and graphic profile become blurred and the resolution of the images has not realized vector control and the amplification effect is very poor and it needs image Mosaic algorithm control.



Fig. 4. Definition in image magnification process after vector field algorithm processing

Compared with Figure 3, image magnification does not appear image resolution in the image vector field Mosaic algorithm mode in the process of magnification. In the amplification process, the image resolution vector has been controlled effectively and the amplification effect is good which verify the effectiveness of the vector field image processing algorithm. It is shown in Figure 4.

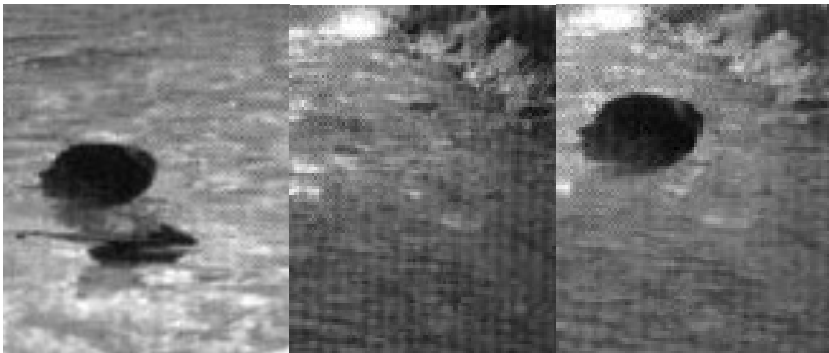


Fig. 5. Vector field synthetic image Mosaic algorithm

Figure 5 is MATLAB image vector field algorithm Mosaic synthetic images. From the image we can see that after the image synthesis, the two kinds of colors of the image has been organic synthesized and synthetic degree is better and the outline of the image has no distortion which proof that the feasibility and effectiveness of image Mosaic technology.

4 The Deficiency of Vector Field Algorithm

We can see the effectiveness of vector field image Mosaic algorithm through the establishment of image Mosaic vector field mathematical model and MATLAB simulation. But there are also a lot of insufficient places such as color image processing in image magnification process; hue saturation processing in image synthesis process; the speed of the computer algorithm vector field.

(1) Color and saturation

In the image magnification and synthetic process, resolution and color contact ratio have been transformed which can ensure there has no distortion situation. But there has the shortage of place of color and tonal processing. Color motif and color saturation not have been transformed which will caused the lost and tonal unsaturated of color.

(2) The speed of the algorithm

With the computer's hardware condition, in order to guarantee algorithm can realized in general computer we must reduce the complexity of the program. So we need to design and optimize algorithm to ensure the stability and rapidity of vector field image Mosaic algorithm.

5 Conclusion

This article established the mathematical model of image Mosaic algorithm through the wizard vector field algorithm and got the 3 d projection coordinate and other important data algorithm through the vector coordinate transformation. This paper done the simulation experiment of image Mosaic algorithm by using MATLAB simulation module and obtained the image magnification and synthesis results. Through the simulation results we found that the common images and images after vector field algorithm Mosaic processing are quite different. Vector field algorithm can guaranteed the image Mosaic process definition have not distortion. In the image synthesis Mosaic simulation we can see that the contact ratio of image through the vector field image Mosaic algorithm is very good. Image Mosaic technology has been widely used in engineering fields. So the development of practical, fast and reliable algorithm is the focus of future researches.

Acknowledgement. The author is much appreciated the financial support for this research from the scientific research project of the Education Department of Guangxi Province with the project number 201106LX578 and the project name Research on Fine Management of Digital Campus Information Based on Data Mining.

References

1. Tolga, T., Pavel, K., Bradley, C.G.: Automatic Mosaicking and Volume Assembly for High-throughput Serial-section Transmission Electron Microscopy. *Journal of Neuroscience Methods* 193(1), 132–144 (2010)

2. Dana, P., Silviu, P., Marc, F.: Experiences in Building a Grid-based Platform to Serve Earth Observation Training Activities. *Computer Standards & Interfaces* 34(6), 493–508 (2012)
3. Nuno, M., Fernando, P.: Automatic Creation and Evaluation of MPEG-7 Compliant Summary Descriptions for Generic Audiovisual Content. *Signal Processing: Image Communication* 23(8), 581–598 (2011)
4. Haipeng, C., Xuanjing, S., Xiaofei, L., Yushan, J.: Bionic Mosaic Method of Panoramic Image Based on Compound Eye of Fly. *Journal of Bionic Engineering* 8(4), 440–448 (2011)
5. Paul, B., Dan, X.: Complex Wavelet-based Image Mosaics Using Edge-preserving Visual Perception Modeling. *Computers & Graphics* 23(3), 309–321 (2011)
6. Dae-Woong, K., Ki-Sang, H.: Practical Background Estimation for Mosaic Blending with Patch-based Markov Random Fields. *Pattern Recognition* 41(7), 2145–2155 (2011)
7. Ali, O., Esra, I.: Marble Mosaic Tiling Automation with a Four Degrees of Freedom Cartesian Robot. *Robotics and Computer-Integrated Manufacturing* 25(3), 589–596 (2011)
8. Helmer, E.H., Ruzycski, T.S., Wunderle Jr., J.M.: Mapping Tropical Dry Forest Height, Foliage Height Profiles and Disturbance Type and Age with a Time Series of Cloud-cleared Landsat and ALI Image Mosaics to Characterize Avian Habitat. *Remote Sensing of Environment* 114(11), 2457–2473 (2010)
9. Nuno, R.G., José, S.V.: Trajectory Reconstruction with Uncertainty Estimation Using Mosaic Registration. *Robotics and Autonomous Systems* 35(3-4), 163–177 (2010)

Conditional Value at Risk Methodology under Fuzzy-Stochastic Approach

Shao-fang Tang¹ and Ying-yu He²

¹ Department of Mathematics, Hangzhou Normal University Qianjiang College,
Hangzhou, China 310036

² Department of Mathematics, Hangzhou Normal University, Hangzhou, China 310036
{tsf_hs, hyy122773}@126.com

Abstract. This paper describes methodology of dealing with financial modeling under uncertainty with risk and vagueness aspects. An approach to modeling risk by the Conditional Value at Risk methodology under imprecise and soft Conditions is solved. It is supposed that the input data and problem conditions are difficult to determine as real number or as some precise distribution function. Thus, vagueness is modeled through the fuzzy numbers of linear T-number type. The combination of risk and vagueness is solved by fuzzy-stochastic methodology. Illustrative example is introduced.

Keywords: Conditional Value at Risk, Uncertainty modeling, Fuzzy sets, Portfolio.

1 Introduction

The major difficulty faced by operation researchers in modeling the problem of portfolio selection is that it is based on the perception of risk by an investor, which will vary because different people have different beliefs about the future performance of various assets. Besides, in real world problems, one is faced with imperfect information (data) and hence must deal with uncertain, imprecise, and vague data. In modeling and analyzing problems of this type, earlier works in finance tended to equate all aspects of imperfect information with uncertainty of a random character. Thus, a multitude of probabilistic models was proposed. This was also the case with the use of modeling in finance. However, no simple and adequate methods for handling imprecise data or vague information, which may stem, for example, from the use of natural language and subjective statements, were available until Zadeh proposed fuzzy sets theory [4]. And indeed, financial modeling has been one of the areas to which fuzzy sets theory has been applied [9].

Typical feature of financial environment is uncertainty. This term is understood mostly as risk uncertainty (probability, stochastic) and is modeled by stochastic apparatus. However, the term uncertainty has the second aspect — the vagueness (sometimes called imprecision, non-preciseness, ambiguity) which is often neglected and could be modeled by fuzzy methodology. In this respect it is apparent that the general

term uncertainty includes two aspects: risk (stochastic) and vagueness (fuzzy) ones. These terms will be used in the paper.

Distinguishing of risk and vagueness confirms a discussion in financial decision making for several years [8]. What are the impacts of uncertainty on decision making? Interesting characteristics are described in Olsen and Throuhton (2000) here vagueness is called ambiguity. First, decision makers are ambiguity averse in general. Second, ambiguity causes more weight to be placed on negative information. Third, risk aversion and ambiguity aversion have not been seen to be highly correlated.

Recently an application of fuzzy-stochastic methodology in financial modeled is being extensively studied now. Valuation of financial instruments where basic assumptions of Von Neumann and Morgenstern expected utility theory are not fulfilled, mainly the assumption of sub-additivity. Particularly, a probability measure is substituted and generalized by a fuzzy measure [12][15]. Forecasting financial characteristics are modeled by robust fuzzy methodology or by neuro-fuzzy network [14]. Multiple criteria evaluation of corporate financial level or credit worthiness with soft aspects and applying fuzzy aggregate operators in comparing with traditional expected utility criterion, Choquet integral, Sugeno integral or order weighting average operators (OWA) [16]. Financial modeling under soft input data and ill-structured decision-making circumstances [13][16].

In the background of the financial globalization, financial risks management is a hot topic for recent years. Risks measurement is the core of effective risks management. In this field the recent developments of financial models are full of successful applications of stochastic models. Now these models are undoubtedly considered as basic tools for financial modeling. One of useful applications is the value at risk (VaR) methodology, which is now a financial standard for modeling the market and credit risk [6](Duffie and Pan, 1997). However, as a risk measure, there is an excellent description of its limitation in Jorion [7]. Similar opinion is presented in Pisoult [10]. The VaR methodology does not provide a measure of worst absolute loss, only some confidence level (risk of exceeding). It lacks sub-additivity and convexity [3]. So the VaR of the combination of two portfolios can be greater than the sum of VaR of the individual portfolios. In addition, the problem of minimizing VaR of a portfolio can have multiple local minimizes. As a result of these drawbacks associated with VaR, the class of coherent risk measures was axiomatized and popularized by Artzner et al. [3] and Delbaen [5]. Of particular interest within this class of risk measures is the conditional value-at-risk (CVaR) risk measure [11], which has received considerable attention within the mathematical finance community.

It is supposed that in conditional value at risk methodology decision-makers attitude is exactly stated and quality (precise) data are at disposal. These assumptions are often fulfilled and stochastic value at risk methodology is sufficient and effective approach for estimation the risk. However, complexity and flexibility of financial world show that there are several situations in which described (goals, criteria and constraints) are not determined precisely and sometimes quality-input data are not at disposal (imprecise input data or distribution functions). Such circumstances exist if

financial system is non-stable, after crisis, in transition economies etc. For modeling such situations, the fuzzy-stochastic approach has been developed and could be applied.

The basic intention of the paper is to show a possibility and procedure for dealing with softly formulated financial problem applying fuzzy-stochastic methodology. The problem on CVaR methodology basis will be described, however, under soft conditions, it means that conditions of decision making are not sharply defined and input data are introduced only vaguely. Such problem could be modeled by the fuzzy-stochastic methodology, it means by a combination of CVaR methodology (risk aspect) and fuzzy methodology (Vagueness aspect).

The paper is organized as follows: in Section 2 we introduce the general concept of conditional value-at-risk, Section 3 is devoted to fuzzy-stochastic methodology description and illustrative example is presented in Section 4.

2 Characterisation of Conditional Value at Risk Methodology

One of the main objectives of risk management is to evaluate and improve the performance of financial organizations in light of the risks taken to achieve profits. A current standard benchmark for firm-wide measures of risk is value at risk (VaR). For a given time horizon T and confidence level β , the value at risk of a portfolio is the loss in the portfolios market value over the time horizon T that is exceeded with probability $1 - \beta$. However, as a risk measure, VaR has recognized limitations. An alternative risk measure to VaR is conditional value at risk (CVaR), which is also known as mean excess loss, mean shortfall or tail VaR. For a given time horizon T and confidence level β , CVaR is the conditional expectation of the loss above VaR for the time horizon T and the confidence level β .

For a given time horizon T , let $f(x, S)$ denote the loss of a portfolio with decision variable $x \in \mathbb{R}^n$ and random variable $S \in \mathbb{R}^d$ denote the value of underlying risk factors at T . Without loss of generality, we assume that the random variable $S \in \mathbb{R}^d$ has a probability density $p(S)$. For a given portfolio x , the probability of the loss not exceeding a threshold α is given by the cumulative distribution function

$$\Psi(x, \alpha) = \int_{f(x, S) \leq \alpha} p(S) dS.$$

When the probability distribution for the loss has no jumps, $\Psi(x, \alpha)$ is everywhere continuous with respect to α .

VaR associated with a portfolio x , for a specified confidence level β and time horizon T , is given by

$$\alpha_\beta(x) = \inf\{\alpha \in \mathbb{R} : \Psi(x, \alpha) \geq \beta\}.$$

Note that under the assumption that $\Psi(x, \alpha)$ is everywhere continuous there exists α (possibly not unique) such that $\Psi(x, \alpha) = \beta$.

Define $[f(x, S) - \alpha]^+$ as

$$[f(x, S) - \alpha]^+ = \begin{cases} f(x, S) - \alpha, & \text{if } f(x, S) - \alpha > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The risk measure CVaR, $\phi(\beta, x)$, is defined as [11]

$$\phi(\beta, x) = \inf_{\alpha} (\alpha + (1 - \beta)^{-1} E([f(x, S) - \alpha]^+)).$$

When the loss distribution has no jumps, CVaR is the conditional expectation of the loss, given that the loss is $\alpha_{\beta}(x)$ or greater, and is given by

$$\phi(\beta, x) = (1 - \beta)^{-1} \int_{f(x, S) \geq \alpha_{\beta}(x)} [f(x, S) - \alpha]^+ p(S) dS.$$

Now consider above formulation based on the standard deviation of normal distributions. Then

$$\begin{aligned} \phi(\beta, x) &= \frac{\varphi(F^{-1}(\beta))}{1 - \beta} \sigma(r_x) - E(r_x) \\ &= \frac{\varphi(F^{-1}(\beta))}{1 - \beta} \sqrt{x^T V x} - E(r_x), \end{aligned}$$

where $\varphi(\cdot)$ is the probability density function of the standard normal distribution, r_x is a portfolio return, $\sigma(r_x)$ is the standard deviation of r_x , F is the cumulative normal distribution function, V is covariance matrix of financial instruments between asset returns and basic random factors returns. In short-term financial decision making it is suppose that mean value $E(r_x)$ is equal to zero, in this way,

$$\phi(\beta, x) = \frac{\varphi(F^{-1}(\beta))}{1 - \beta} \sqrt{x^T V x}. \tag{1}$$

This is well-known simplified formula of CVaR methodology, as well.

Having in mind the basic goal of the paper is to show and apply a fuzzy-stochastic methodology; therefore it is necessary to provide a description of CVaR modeling circumstances. We shall introduce assumptions and input parameters that could be difficult to determine precisely either real numbers or distribution functions. This aspect also implies from the decision-making process conditions. The reasons why the CVaR methodology parameters could be understood as uncertain imply from the input data uncertainty and the approximate procedures used. There is an excellent

description of limitation of value at risk methodology in Jorion [7]. Similar opinion is presented in Pisoult [10].

We can summarize the findings in the following way and stress these aspects, which are useful in modeling vague parameters in CVaR determination.

(1) Probability level because of difficulty to state parameter β precisely; knowledge of sensitivity of the CVaR level (β) is useful. (2) Holding period (T) should be suitable to consider in some intervals, it means vaguely. (3) A portfolio composition vector (x) because assumption that a portfolio composition does not remain unchanged throughout the holding period is more realistic; because in a financial planning process a portfolio composition is not often known precisely in the decision-making moment. (4) Multivariate normal distribution approximation is often used. (5) Historical data are used for forecasting and it is assumed that the situation of financial instruments in the future will be similar to the past. Hence, the forecasting is more risky and vague. (6) There are problems with input quality of data, data frequency (expected returns, correlation and volatility) and stochastic validity.

3 Fuzzy-Stochastic Model under Normal Fuzzy Sets

Apparently, there are two aspects presented in decision-making process, risk and vagueness. The risk is understood to be a stochastic aspect and we can suppose that a distribution function is known. On the other hand vagueness means that the location and shape of the distribution are opened to question and so it is a function of confidence degree or the weight attached to the probability judgment. It is the basic idea and intention how to deal with combination of risk and vagueness. Other parameters are possible to define vaguely, as well. The risk is modeled by random methodology and fuzzy apparatus is used for vagueness modeling. Thus one of the suitable approaches for solving this problem is to apply a fuzzy-stochastic methodology and to create a fuzzy-stochastic model.

There are several basic fuzzy-stochastic elements, which are very important from an application point of view.

Let R be the set of all real number endowed with a usual topology. Then a fuzzy subset \tilde{A} of R is defined by its membership function $\tilde{A} : R \rightarrow [0,1]$. We denote by $\tilde{A}_\alpha = \{x : \tilde{A}(x) \geq \alpha\}$ the α -level set of \tilde{A} for $\alpha \in [0,1]$. The 0-level set \tilde{A}_0 of \tilde{A} is defined by the closure of the set $\{x : \tilde{A}(x) > 0\}$. Now \tilde{A} is called a normal fuzzy set if there exists an x such that $\tilde{A}(x) = 1$, and \tilde{A} is called a convex fuzzy set if $\tilde{A}(\lambda x + (1 - \lambda)y) \geq \min\{\tilde{A}(x), \tilde{A}(y)\}$ for $\lambda \in [0,1]$; that is, \tilde{A} is a quasi-concave function.

Let f be a real-valued function defined on R . Then f is said to be upper semi continuous if $\{x : f(x) > \alpha\}$ is a closed set for each α . Or, equivalently, f is upper semi continuous at y if and only if, for each $\varepsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta$ implies $f(x) < f(y) + \varepsilon$.

Definition 3.1 Let \tilde{A} be a fuzzy subset of \mathbb{R} , then \tilde{A} is called a fuzzy number if the following conditions are satisfied:

- (i) \tilde{A} is a normal and convex fuzzy set;
- (ii) its membership function $\tilde{A}(x)$ is upper semi continuous;
- (iii) the 0-level set \tilde{A}_0 is bounded.

The above condition (iii) also says that each α -level set \tilde{A}_α is bounded for $\alpha \in [0,1]$. From Zadeh [4], \tilde{A} is a convex fuzzy set if and only if each of the α -level set \tilde{A}_α is a convex set. Therefore if \tilde{A} is a fuzzy number, then the α -level set \tilde{A}_α is a compact (closed and bounded in \mathbb{R}) and convex set; \tilde{A}_α is a closed interval. Then the α -level set \tilde{A}_α is denoted by $\tilde{A}_\alpha = [A_\alpha^L, A_\alpha^U]$. The following proposition is useful for further discussions.

Proposition 3.1 (Resolution identity [4]). Let \tilde{A} be a fuzzy set and $\tilde{A}_\alpha = \{x : \tilde{A}(x) \geq \alpha\}$, then

$$\tilde{A}(x) = \sup_{\alpha \in [0,1]} \alpha \cdot 1_{\tilde{A}_\alpha}(x). \tag{2}$$

Where $1_{\tilde{A}_\alpha}$ is an indicator function of set \tilde{A}_α .

We denote by \mathbf{F} the set of all fuzzy subsets of \mathbb{R} . Let $f(x_1, x_2, \dots, x_n)$ be a real-valued function from \mathbb{R}^n into \mathbb{R} and $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ be n fuzzy sets of \mathbb{R} . By the extension principle in Zadeh [4], we can induce a fuzzy-valued function $\tilde{f} : \mathbf{F}^n \rightarrow \mathbf{F}$ according to the real-valued function $f(x_1, x_2, \dots, x_n)$. In other words, $\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)$ is a fuzzy subset of \mathbb{R} . Then the membership function of $\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)$ is defined by

$$\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)(y) = \sup_{(x_1, x_2, \dots, x_n) \in f^{-1}(y)} \min\{\tilde{A}_1(x_1), \tilde{A}_2(x_2), \dots, \tilde{A}_n(x_n)\}. \tag{3}$$

Proposition 3.2 [4] Let S be a compact set in \mathbb{R}^n . If f is upper semi continuous on S then f attains maximum over S and if f is lower semi continuous on S then f attains minimum over S .

Proposition 3.3 Let $f(x_1, x_2, \dots, x_n)$ be a real-valued function defined on \mathbb{R}^n and $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ be n fuzzy sets of \mathbb{R} . Let $\tilde{f} : \mathbf{F}^n \rightarrow \mathbf{F}$ be a fuzzy-valued function induced by $f(x_1, x_2, \dots, x_n)$ via the extension principle defined in (2). Suppose that each membership function \tilde{A}_i is upper semi continuous

on \mathbb{R} for $i = 1, 2, \dots, n$ and each $\{(x_1, x_2, \dots, x_n) : y = f(x_1, x_2, \dots, x_n)\}$ is a compact subset of \mathbb{R}^n (it will be a closed and bounded set in \mathbb{R}^n) for y in the range of f . Then the α -level set of $\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)$ is

$$[\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)]_\alpha = \{f(x_1, x_2, \dots, x_n) : x_1 \in [\tilde{A}_1]_\alpha, x_2 \in [\tilde{A}_2]_\alpha, \dots, x_n \in [\tilde{A}_n]_\alpha\}$$

Especially, we use above proposition to fuzzy numbers, then the following results are not hard to prove.

Proposition 3.4 Let $f(x_1, x_2, \dots, x_n)$ be a continuous real-valued function defined on \mathbb{R}^n and $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ be n fuzzy sets of \mathbb{R} . Let $\tilde{f} : \mathbb{F}^n \rightarrow \mathbb{F}$ be a fuzzy-valued function induced by $f(x_1, x_2, \dots, x_n)$ via the extension principle defined in (3). Suppose that each $\{(x_1, x_2, \dots, x_n) : y = f(x_1, x_2, \dots, x_n)\}$ is a compact subset of \mathbb{R}^n for y in the range of f . Then $\tilde{f}(\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n)$ is a fuzzy number and its the α -level set is

$$[\tilde{f}(\tilde{A}_1, \dots, \tilde{A}_n)]_\alpha = \{f(x_1, \dots, x_n) : x_1 \in [\tilde{A}_1]_\alpha, \dots, x_n \in [\tilde{A}_n]_\alpha\}$$

$$= \{f(x_1, \dots, x_n) : [A_1]_\alpha^L \leq x_1 \leq [A_1]_\alpha^U, \dots, [A_n]_\alpha^L \leq x_n \leq [A_n]_\alpha^U\}$$

where $[\tilde{A}_i]_\alpha = [[A_i]_\alpha^L, [A_i]_\alpha^U] (i = 1, 2, \dots, n)$.

Definition 3.2 [2] Let (Ω, A, P) be a probability measure space. A mapping $\tilde{A} : \Omega \rightarrow \mathbb{F}$ is called a fuzzy random variable on (Ω, A) , if for any $\alpha \in [0, 1]$

$$\tilde{A}_\alpha = \{x : \tilde{A}(\omega) \geq \alpha\} = [\tilde{A}_\alpha^L, \tilde{A}_\alpha^U]$$

is a random interval, that is $\tilde{A}_\alpha^L, \tilde{A}_\alpha^U$ are two random variables on (Ω, A) .

4 Illustrative Example of Fuzzy-Stochastic CVaR Calculation

Since the purpose of the paper is to show an application possibility of the fuzzy-stochastic approach for stating minimizing CVaR under uncertain (risk, vagueness) circumstances, the following examples describes a possibility of estimation the CVaR based on fuzzy-stochastic conditions. Uncertain CVaR calculation is based on Eq. (1), which is supported by Eq. (2) and is formulated as resolution (decomposition) principle function and by analogy it is solved by several non-linear mathematical problems.

Problem

$$\begin{aligned} \max(\min) \quad \text{CVaR} &\equiv \text{CVaR}_\alpha^L, \text{CVaR}_\alpha^U \quad \text{for } \alpha \in [0, 1] \\ \text{s.t.} \quad \text{CVaR} &= \frac{\varphi[F^{-1}(\beta)]}{1 - \beta} \sqrt{x^T V x} \\ V &\in [\tilde{V}_\alpha^L, \tilde{V}_\alpha^U], x \in [\tilde{x}_\alpha^L, \tilde{x}_\alpha^U], T \in [\tilde{T}_\alpha^L, \tilde{T}_\alpha^U], \beta \in [\tilde{\beta}_\alpha^L, \tilde{\beta}_\alpha^U]. \end{aligned}$$

As there are infinite ways to characterize fuzziness, there are infinite ways to graphically depict the membership and to generate the data. Normally, experts should be able to offer decision makers or investors information regarding the measure of fuzziness. In this context, fuzzy input data are given by the following way. The expiry fuzzy period T and probability fuzzy level β are given subjectively by a financial analyst. The fuzzy covariance matrix V is determined similarly by methodology described in Zmeskal [16] applying the statistical error due to the simplified formula $\hat{\sigma}^2 = N(\sigma^2, 2\sigma^4 / (T - 1))$. Simultaneously quintals of statistical error distribution determine a shape of the linear T-number.

Definition 5.1 The linear T-number is defined by:

$$\tilde{s}(x) = \begin{cases} \frac{x - (s^L - s^a)}{s^a}, & \text{if } s^L - s^a < x < s^L, \\ 1, & \text{if } s^L \leq x \leq s^U, \\ \frac{(s^U + s^b) - x}{s^b}, & \text{if } s^U < x < s^U + s^b, \\ 0, & \text{otherwise.} \end{cases}$$

In this paper, we consider that $s^L = 40\%$, $s^U = 60\%$, $s^L - s^a = 5\%$ and $s^U + s^b = 95\%$. An analyst can introduce some input data by linguistic term, as well. Input data matrices of linear T-number have the following form:

$$\tilde{s} = \begin{bmatrix} s^L & s^U \\ s^a & s^b \end{bmatrix}$$

We assume to have four assets: home stock, foreign stock, bond 2 year, bond 3 year. For the sake of simplicity non-precise input data are introduced as the linear T-numbers. It is not difficult to solve the problem with piecewise fuzzy numbers, of course. Now let us show the fuzzy covariance matrix in Table 1, the fuzzy portfolio composition vector \tilde{x} in Table 2, and the fuzzy numbers $\tilde{\beta}$ and \tilde{T} have the following form:

$$\tilde{\beta} = \begin{bmatrix} 0.949 & 0.950 \\ 0.006 & 0.003 \end{bmatrix} \quad \tilde{T} = \begin{bmatrix} 0.300 & 0.320 \\ 0.010 & 0.020 \end{bmatrix}$$

Calculated uncertain CVaR for α -level sets with numbers 0, 0.25, 0.5, 0.75 and 1.0 are clear from Table 3. It is obvious that the minimum loss will range from 10.26 to 22.16 for α -level 0 and from 14.59 to 16.31 for α -level 1.00 under the written prepositions. Thus result might be understood as the fuzzy-probability decision making space and information about sensitivity of the CVaR value on input data vagueness too. For the sake of comparison, crisp-stochastic solution is presented too; CVaR value is 14.61. It shows that in this example the result is not in the centre of T-number, and potential loss is underestimated.

Table 1. The Fuzzy Covariance Matrix

		A	B	C	D	
Home stock	A	0.02250	0.02625	-0.00540	-0.00525	s^L
Foreign stock	B	0.02625	0.06250	-0.00900	-0.00875	
Bond 2 year	C	-0.00540	-0.00900	0.00810	0.00567	
Bond 3 year	D	-0.00525	-0.00875	0.00567	0.00490	
Home stock	A	0.02273	0.0256	-0.00497	-0.00520	s^U
Foreign stock	B	0.02651	0.06563	-0.00828	-0.00866	
Bond 2 year	C	-0.00535	-0.00855	0.00875	0.00573	
Bond 3 year	D	-0.00520	-0.00875	0.00612	0.00495	
Home stock	A	0.00225	0.00525	0.00108	0.00053	s^a
Foreign stock	B	0.00263	0.01250	0.00180	0.00088	
Bond 2 year	C	0.00054	0.00180	0.00162	0.00057	
Bond 3 year	D	0.00053	0.00175	0.00113	0.00049	
Home stock	A	0.00068	0.00110	0.00015	0.00010	s^b
Foreign stock	B	0.00080	0.00263	0.00025	0.00017	
Bond 2 year	C	0.00016	0.00034	0.00026	0.00011	
Bond 3 year	D	0.00016	0.00033	0.00018	0.0010	

Table 2. Vector \tilde{x}

	A	B	C	D	sum
s^L	20.0	40.0	30.0	10.0	100.0
s^U	21.0	41.0	31.0	11.0	104.0
s^a	2.0	8.0	3.0	2.0	15.0
s^b	2.1	12.3	3.1	2.2	29.7

Table 3. CvaR in α -level

	α	min	max		α	min	max
Fuzzy	0.00	10.26	22.16	Fuzzy	0.75	13.43	17.67
Fuzzy	0.25	11.27	20.61	Fuzzy	1.00	14.59	16.31
Fuzzy	0.50	22.33	19.12	Crisp	1.00	14.61	14.61

5 Conclusion

The methodology of Conditional Value at Risk is now under deep research and application verification. One of the aspects studies is a possibility to introducing various preciousness of datum and applying an approximate and simplified methodology. Unfortunately, in many situations it is very difficult to determine these inputs because the conditions and circumstances of financial decision making are often uncertain. On the other hand it is useful to use the vague distribution functions. The application of the fuzzy-stochastic methodology with the normal fuzzy numbers is one of practical possibilities how to solve this problem. The application of the methodology was

presented for the sake of simplicity on the analytical method. The described approach might be seen as the generalized sensitivity analysis of calculation the CVaR value under imprecise input data and decision-making conditions. The fuzzy-stochastic approach described is suitable for ill structures and non-stable situations with vaguely stated input data and probability distributions. On the other hand the crisp-stochastic approach could also be applied if well structures problem with precisely data and probability exists. These two approaches are not in contradiction, but we can consider them to be complement.

References

1. Alexander, S., Coleman, T.F., Li, Y.: Minimizing CVaR and VaR for a portfolio of derivatives. *Journal of Banking and Finance* 30, 583–605 (2006)
2. Ammar, E.E.: On solutions of fuzzy random multiobjective quadratic programming with applications in portfolio problem. *Information Sciences* 178, 468–484 (2008)
3. Artzner, P., Eber, F., Eber, J.M., Heath, D.: Coherent measures of risk. *Mathematical Finance* 9, 203–228 (1999)
4. Bellman, R.E., Zadeh, L.A.: Decision-making in a fuzzy environment. *Management Science* 17, 141–164 (1970)
5. Delbaen, F.: Coherent measures of risk on general probability spaces. In: Sandmann, K., Schonbucher, P.J. (eds.) *Advances in Finance and Stochastics, Essays in Honor of Dieter Sondermann*, pp. 1–37. Springer, Berlin (2002)
6. Duffie, D., Pan, J.: An overview of value at risk. *Journal of Derivatives*, 7–49 (Spring 1997)
7. Jorion, P.: *Value at Risk*. Mc-Graw Hill
8. Olsen, R.A., Throuhton, G.H.: Are risk premium anomalies caused by ambiguity? *Financial Analyst Journal* (March/April 2000)
9. Ostermark, R.: Fuzzy linear constrains in the capital asset pricing model. *Fuzzy Sets and Systems* 30, 93–102 (1989)
10. Pisoult, E.: Quantifying the risk of trading. In: Dempster, M.A.H. (ed.) *Risk Management: Value at Risk and Beyond*. Cambridge University Press
11. Rockafellar, R.T., Uryasev, S.P.: Optimization of conditional value-at-risk. *J. Risk* 2, 21–41 (2000)
12. Simonelli, M.R.: Fuzziness in valuing financial instruments by certainty equivalents. *European Journal of Operational Research* 135, 296–302 (2001)
13. Tanaka, H., Guo, P., Turksen, I.B.: Portfolio selection based on fuzzy probabilities and possibility distributions. *Fuzzy Sets and Systems* 111, 387–397 (2000)
14. Tseng, F.M., Yu, G.H., et al.: Fuzzy ARIMA model for forecasting the foreign exchange market. *Fuzzy Sets and Systems* 118, 9–19 (2001)
15. Young, V.R., Zarihopoulu, T.: Computation of distorted probabilities for diffusion processes via stochastic control methods. *Insurance, Mathematics and Economics* 27, 1–18 (2000)
16. Zmeskal, Z.: Application of the fuzzy-stochastic methodology to appraising the firm value as a European call option. *European Journal of Operational Research* 135(2), 303–310 (2001)

Application of the Fuzzy Set with Three Kinds of Negation FSCOM in the Stock Investment

Jiexin Zhao¹ and Zhenghua Pan²

School of Science, Jiangnan University, Wuxi, China
15161980890@163.com

Abstract. Paper [1] has proposed that negation of the fuzzy knowledge include the contradictory negation, the opposite negation and the medium negation, and defined a new fuzzy set with *contradictory negation*, *opposite negation* and *medium negation* (FSCOM). To show the applicability of FSCOM, it was applied in an example about the stock investment decision-making. An approach to confirm the membership functions of fuzzy sets is presented. Threshold value of range of membership degree is established. Reasoning and realization of fuzzy decision making in the example is discussed based on FSCOM and the fuzzy production rule. It implies that the application of FSCOM in dealing with practical problems with ambiguity and different negations is effective.

Keywords: fuzzy set FSCOM, the stock investment, fuzzy decision, threshold λ .

1 Introduction

The negation of the fuzzy information is an important and special concept in fuzzy information processing [2]. Referring to the cognition and processing of negative information, there are some theories and methods. For example, FS (Zadeh's fuzzy sets), IFS (intuitionistic fuzzy sets), Interval valued fuzzy sets, VS (Vague Sets), RS (rough sets), etc. But for the cognition and processing of negation, such as Classic set, it has a kind of negation. So, it is hard to do the processing of distinguish, expression, and computing for various negations of fuzzy information which exists in objective world widely [3]. For this, some scholars proposed concepts and methods of processing different negative information [4-7]. In 2011, Pan zhenghua defined FScom which is a new fuzzy set with contradictory negation, opposite negation and medium negation [1]. He has also researched the properties of FScom and its practical application in fuzzy financial decision-making [2].

Based on the theory of paper [1], the application of FScom has been researched further. By the example of fuzzy decision-making about the stock investment and FScom, an approach to confirm the membership function of fuzzy sets will be introduced. Threshold value of range of membership degree will be established. This paper also discusses reasoning and realization of fuzzy decision making in the example. That shows the applicability of FScom in dealing with the objective actual problems.

2 Basis

Definition 1. Let U be domain. Mapping

$$\Psi_A : U \rightarrow [0, 1]$$

confirms a fuzzy subset A on U , i.e. $A \in P(U)$, where mapping Ψ_A is called the membership function of A , $\Psi_A(x)$ is the degree of membership of x on A , for short $A(x)$. $\{A(x) \mid x \in U\}$ for short $A(U)$.

Definition 2. Let $A \in P(U)$, $\lambda \in (0, 1)$.

① Mapping

$$\Psi^{\neg} : A(U) \rightarrow [0, 1]$$

Ψ^{\neg} confirms a fuzzy subset (written as A^{\neg}) on U if $\Psi^{\neg}(A(x)) = 1 - A(x)$, which $A^{\neg}(x) = \Psi^{\neg}(A(x))$. A^{\neg} is called *opposite negation set of A*.

② Mapping

$$\Psi^{\sim} : A(U) \rightarrow [0, 1]$$

Ψ^{\sim} confirms a fuzzy subset (written as A^{\sim}) on U if $A^{\sim}(x) = \Psi^{\sim}(A(x))$, which $A^{\sim}(x) =$

$$\left\{ \begin{array}{ll} \lambda - \frac{2\lambda - 1}{1 - \lambda} (A(x) - \lambda), & \lambda \in [1/2, 1) \text{ and } A(x) \in (\lambda, 1] \quad (1) \\ \lambda - \frac{2\lambda - 1}{1 - \lambda} A(x), & \lambda \in [1/2, 1) \text{ and } A(x) \in [0, 1 - \lambda) \quad (2) \\ 1 - \frac{1 - 2\lambda}{\lambda} A(x) - \lambda, & \lambda \in (0, 1/2] \text{ and } A(x) \in [0, \lambda) \quad (3) \\ 1 - \frac{1 - 2\lambda}{\lambda} (A(x) + \lambda - 1) - \lambda, & \lambda \in (0, 1/2] \text{ and } A(x) \in (1 - \lambda, 1] \quad (4) \\ A(x), & \text{other} \quad (5) \end{array} \right.$$

A^{\sim} is called *medium negation set of A*.

③ Mapping

$$\Psi^{\neg} : A(U) \rightarrow [0, 1]$$

Ψ^{\neg} confirms a fuzzy subset (written as A^{\neg}) on U if $\Psi^{\neg}(A(x)) = \text{Max}(A^{\neg}(x), A^{\sim}(x))$, which $A^{\neg}(x) = \Psi^{\neg}(A(x))$. A^{\neg} is called *contradictory negation set of A*.

The fuzzy subsets on domain U is defined by the definition 1 and definition 2, which is called ‘Fuzzy Set with Contradictory negation, Opposite negation and Medium negation’, for short FScom [1].

3 The Application of FSCom in the Stock Investment

Fuzzy decision-making deals with things in the environment of non-probability and vagueness. Fuzzy set is usually the most suitable method, when people need to construct knowledge model for non-probability and vagueness. Then, how to apply FScom in fuzzy decision-making? We may take the following practical problem as an example.

3.1 An Example

In real life, how to individuals (people or family) invest in the stock market? We take statistics about the stock market comes from Statistical information Wind [8] for analysis (Table 1). We know that it's better when the stock price increase bigger. According to the overseas mature stock markets' tradition and the actual situation of Chinese stock market, most investors consider that stocks with lower than 20 times P/E ratio have investment value. But, it's more dangerous when P/E ratio exceed 30 or 40 times [9]. If P/E ratio is high, the deviation degree from the price and the value is high. Then, the time of taking back the capital is long. In that sense, it is better if P/E ratio is as low as possible. If P/E ratio is high, the investment recovery period is longer, and the investment risk is bigger [10]. Through the above, we set DR (the Decision Rules) of the stock investment:

- If amount of increase is small and P/E ratio is high, investor should not invest the stock.
- If amount of increase is big and P/E ratio is low, investor should invest the stock.
- If amount of increase is average and P/E ratio is low, investor should invest most of the capital in the stock.
- If amount of increase is small and P/E ratio is moderate, or amount of increase is big and P/E ratio is high, investor should invest a small part of the capital in the stock.
- If amount of increase is average and P/E ratio is moderate, or amount of increase is small and P/E ratio is low, investor should invest half of the capital in the stock.

Table 1. Statistics about the *price change (%)* and the *P/E ratio (time)* of the stocks

Security (Abbreviation)	Price change	P/E ratio	Security (Abbreviation)	Price change	P/E ratio
Hareon Solar	203.20	25.93	GOLD MANTIS	64.38	36.63
Sealand Securities	201.73	159.95	Shandong Molong	71.72	39.82
GRANDHOPE BIOTECH	63.07	90.09	Shanghai Construction	75.42	12.12
VVFB	146.57	102.18	CORUN	72.24	699.83
Baotou Rare Earth	143.07	26.25	Jinfeng investment	72.11	20.10
7 Shares	139.36	582.59	JJUGUI	71.71	45.13
ZHONG KE SAN HUAN	117.41	23.55	Guangzhou Pharmaceutical	67.73	55.11
Kile technology	113.87	50.30	JIANFENG GROUP	69.81	13.18
ST Chang Jiu	113.75	527.97	Chongqing industrial	69.36	22.04
Ear HengYun A	111.11	38.46	GuanCheng chase	69.10	8.71
Zhejiang east	180.79	61.57	Suning Universal	68.13	22.62
Rising Nonferrous	92.97	116.22	*ST Garden city	63.28	61.99
St BAOLONG	92.40	2289.70	MTC	67.62	20.05
Tuopai Shede	91.16	39.65	Caticreal estate	67.60	6.69
Kingdream	90.90	82.89	XTC	67.48	33.01
Science City	85.62	195.74	Dulun stake	78.66	156.98
JOIN-IN	83.00	21.58	SongDou shares	82.04	15.66
YONGAN FOREST	69.61	-35.22	DongBao Biology	94.07	91.85

3.2 Representation of Fuzzy Sets and Different Negations in DR and the Example

It is obviously that

- ‘low P/E ratio’, ‘high P/E ratio’, ‘moderate P/E ratio’, ‘small amount of increase’, ‘big amount of increase’, ‘average amount of increase’ are different fuzzy sets.
According to FScom, we need to point out that there are following relationships among these fuzzy sets:
- ‘small amount of increase’ is the opposite negation of ‘big amount of increase’; ‘average amount of increase’ is the medium negation of ‘big amount of increase’; ‘low P/E ratio’ is the opposite negation of ‘high P/E ratio’; ‘moderate P/E ratio’ is the medium negation of ‘high P/E ratio’.
- The existing fuzzy set theory to deal with these fuzzy sets and their relations has limitations. Because they do not distinguish contradictory negation, opposite negation and medium negation, so the process to deal with the problems with fuzziness and different negations is complicated.

Based on FScom, different fuzzy sets in DR can be represented: MUCHincrease denotes ‘big amount of increase’; MUCHincrease¹ denotes ‘small amount of increase’; MUCHincrease~ denotes ‘average amount of increase’; MUCHyield denotes ‘high P/E ratio’; MUCHyield¹ denotes ‘low P/E ratio’; MUCHyield~ denotes ‘moderate P/E ratio’.

And MUCHincrease¹ is the opposite negation of MUCHincrease, MUCHincrease~ is the medium negation of MUCHincrease; Highyield¹ is the opposite negation of Highyield, Highyield~ is the medium negation of Highyield.

In addition, the formal denotations of decision-making behavior are: Stock(no): ‘investor don’t invest the stock’; Stock(yes): ‘investor invest the stock’; Stock(more): ‘investor invest most of the capital in the stock’; Stock(little): ‘investor invest a small part of the capital in the stock’; Stock(half): ‘investor invest half of the capital in the stock’.

Let x be a datum in the table 1, the degree of membership of x on the fuzzy set:

- MUCHincrease(x): denotes degree of membership of x on MUCHincrease;
- MUCHincrease¹(x): denotes degree of membership of x on MUCHincrease¹;
- MUCHincrease~(x): denotes degree of membership of x on MUCHincrease~;
- MUCHyield(x): denotes degree of membership of x on MUCHyield;
- MUCHyield¹(x): denotes degree of membership of x on MUCHyield¹;
- MUCHyield~(x): denotes degree of membership of x on MUCHyield~.

Therefore, the formal representation of decision rules is:

- a) Muchincrease¹(x) \wedge Highyield(x) \rightarrow Stock(no)
- b) Muchincrease(x) \wedge Highyield¹(x) \rightarrow Stock(yes)
- c) Muchincrease~(x) \wedge Highyield¹(x) \rightarrow Stock(more)

- d) $(\text{Muchincrease}^{\bar{1}}(x) \wedge \text{Highyield}^{\bar{1}}(x)) \vee (\text{Muchincrease}(x) \wedge \text{Highyield}(x)) \rightarrow \text{Stock}(\text{little})$
- e) $(\text{Muchincrease}^{\bar{1}}(x) \wedge \text{Highyield}^{\bar{1}}(x)) \vee (\text{Muchincrease}^{\bar{1}}(x) \wedge \text{Highyield}^{\bar{1}}(x)) \rightarrow \text{Stock}(\text{half})$

3.3 The Establishment of Membership Function of Fuzzy Set in DR and the Example

In the example, how to determine the investment strategy in DR? Because the premise of DR is mainly composed by above fuzzy sets, so, the establishment of membership functions of fuzzy sets is the basis of decision-making. Based on FSCOM and true value measurement method of fuzzy predicate [11], we present a new approach to confirm membership functions of different fuzzy sets as follows:

First, let $V = \{V_1, V_2, \dots, V_n\}$ be the set of index need to be analysis when making decision. And let $U = \{U_1, U_2, \dots, U_m\}$ be the set of stocks to be analysis.

Second, let x_{ij} be the index value of the i th indexes of the j th stock. Let $M_i = \max_j \{x_{ij}\}$ be the upper limit on V_i , whereas its lower limit is $L_i = \min_j \{x_{ij}\}$, where $i = 1, 2, \dots, n, j = 1, 2, \dots, m$.

Finally, define H_i as the fuzzy set of V_i , Based on the definition of FSCOM, $H_i^{\bar{1}}$ is the opposite negation of H_i , and $H_i^{\bar{2}}$ is the medium negation of H_i .

Since decision interval of the fuzzy set H_i is defined according to the experience of stock investors and the statistics in the table 1. Therefore, it cannot reflect the actual situation accurately. In order to increase the accuracy of decision-making, we give elasticity values to the upper limit and the lower limit of the decision interval.

As for the definition method of membership function of the fuzzy set, the Euclidean distance of one-dimension space which $d(x, y) = |x - y|$ and the definition of distance ratio function [11] are adopted. The membership function of the fuzzy set can be defined as follows:

Definition 3. Let x be index value of V_i , H_i be the fuzzy set of index V_i , Then membership function of H_i is:

$$H_i(x) = \begin{cases} 0, & x \leq L_i + \varepsilon_{L_i} \\ \frac{d(x, L_i + \varepsilon_{L_i})}{d(L_i + \varepsilon_{L_i}, M_i - \varepsilon_{M_i})}, & L_i + \varepsilon_{L_i} < x < M_i - \varepsilon_{M_i}, i=1, 2, \dots, n \\ 1, & x \geq M_i - \varepsilon_{M_i} \end{cases}$$

Where, M_i is the upper limit of the decision interval on H_i and its elasticity values is ε_{M_i} . L_i is the lower limit of the decision interval on H_i and ε_{L_i} is elasticity values.

Now, the task is to solve the following problem following the above steps. Consider two indexes: amount of increase V_1 and P/E ratio V_2 .

Based on the data in table 1, we can get the result: $M_1 = 203.20\%$, $L_1 = 63.07\%$. Because when the earning per share is negative, PE ratio is meaningless. From its essence sense, positive P/E ratio is static pay-back period, and negative P/E ratio is absolutely pointless [12]. From [9] and [10], we can conclude that P/E ratio is thought to be low when it is approximately less than 40 times. P/E ratio is thought to be high when it is approximately higher than 100 times. The stock with too high P/E ratio has higher risk, and the earning cycle is longer. So, the stock has P/E ratio that is negative or more than 200 times is just beyond the decision-making scope. At this point, $M_2 = 195.74$ times and $L_2 = 6.69$ times can be got.

In order to increase the accuracy of decision-making, take elasticity values (table 2).

Table 2. Elasticity Values (ϵ) of upper and lower limit

Elasticity Values	Amount of Increase	P/E ratio
Upper Limit	$\epsilon_{m_1} = \pm 1$	$\epsilon_{m_2} = \pm 1$
Lower Limit	$\epsilon_{l_1} = \pm 1$	$\epsilon_{l_2} = \pm 1$

Thus, combing the definition 3, membership function of Muchincrease is:

$$\text{Muchincrease}(x) = \begin{cases} 0, & x \leq 64.07 \\ \frac{d(x,64.07)}{d(64.07,202.20)}, & 64.07 < x < 202.20 \\ 1, & x \geq 202.20 \end{cases} \quad (6)$$

According to definition 2, membership function of Muchincrease¹ is:

$$\text{Muchincrease}^1(x) = 1 - \text{Muchincrease}(x). \quad (7)$$

In the same way, membership function of Highyield and Highyield¹ are:

$$\text{Highyield}(x) = \begin{cases} 0, & x \leq 7.7 \\ \frac{d(x,7.69)}{d(7.69,194.74)}, & 7.69 < x < 194.74 \\ 1, & x \geq 194.74 \end{cases} \quad (8)$$

$$\text{Highyield}^1(x) = 1 - \text{Highyield}(x). \quad (9)$$

The next step is determining membership functions of MUCHincrease[~] and Highyield[~]. By (1) – (5), we can determine membership functions of the fuzzy sets: MUCHincrease[~], Highyield[~]. However, the formulas relates to the establishment of λ . Therefore, the following provides a solution of establishing λ .

3.4 Establishment Method and Meaning of λ for FSCOM in the Example

In need of analysis, we take the 5th amount of increase as upper limit of it and take the 9th from the bottom of amount of increase as lower limit of it. According to 3.3, it is

known that the P/E ratio is thought to be low when below 40 times. P/E ratio is thought to be high when above 100 times. The stock with negative P/E ratio or higher than 200 times is just beyond the scope of decision-making. So, following table can be confirmed (table 3):

Table 3. The Decision Intervals (DI) on Amount of Increase and P/E ratio

Fuzzy Set	MUCHincrease	MUCHincrease [¬]	MUCHyield	MUCHyield [¬]
DI	[143.07, 203.20]	[63.07, 68.13]	[102.18, 195.74]	[6.69, 39.82]

Combining elasticity values in the table 2 and the decision intervals in the table 3, it can provide each index data flexible intervals to reflect its accuracy: i Flexible intervals of upper and lower limit on MUCHincrease are respectively [202.20, 204.20], [142.07, 144.07]. 142.07% is the lower limit on MUCHincrease. ii Flexible intervals of upper and lower limit on Muchincrease[¬] are respectively [67.13, 69.13], [62.07, 64.07]. 69.13% is the upper limit on Muchincrease[¬].iii Flexible intervals of upper and lower limit on Highyield are respectively [194.74, 196.74], [101.18, 103.18]. 101.18 times is the lower limit on Highyield.iv Flexible intervals of upper and lower limit on Highyield[¬] are respectively [38.82, 40.82], [5.69, 7.69]. 40.82 times is the upper limit on Highyield[¬].

From (6) and (7), Muchincrease(142.07)≈ 0.565; Muchincrease(69.13)≈ 0.037; Muchincrease[¬](69.13) = 0.963. Due to 142.07% and 69.13% were from two opposite fuzzy sets: Muchincrease and Muchincrease[¬] respectively, and 142.07% is the upper limit on Muchincrease, 69.13% is the lower limit of Muchincrease[¬]. So, degrees of membership of them on Muchincrease ought to theoretically satisfy: Muchincrease(142.07) =1, and Muchincrease[¬](69.13) =1. However, because of data insufficiency and data distortion, the equation is not set up. For this, we quote the method in [13]. That is taking $\frac{1}{2}(\text{Muchincrease}(142.07) + \text{Muchincrease}^{\neg}(69.13)) = 0.764$ as an equilibrium amount, which denoted a limit of big amount of increase. It means that if Muchincrease(x) ≥ 0.764 such that for an arbitrary value x%, x belongs to Muchincrease. Thus, we regard 0.764 as threshold for the fuzzy set ‘big amount of increase’.

Similarly, by (8) and (9), the equilibrium amount of Highyield(x) is $\frac{1}{2}(\text{Highyield}(101.18) + \text{Highyield}^{\neg}(40.82)) = 0.662$. With this, λ on Muchincrease and Highyield can be established (Table 4).

Table 4. λ on Muchincrease and Highyield

Index	Threshold value λ
Increase	λ ₁ = 0.764
P/E ratio	λ ₂ = 0.662

Based on FSCom, threshold value λ has the following meaning:

- 1) If Muchincrease(x) ≥ λ (or Highyield(y) ≥ λ), the stock has big amount of increase (or high P/E ratio).
- 2) If Muchincrease(x) ≤ 1-λ (or Highyield(y) ≤ 1-λ), the stock has small amount of increase (or low P/E ratio).

3) If $1-\lambda < \text{Muchincrease}(x)$ (or $\text{Highyield}(y) < \lambda$, the stock has average amount of increase (or moderate P/E ratio) (x is the value of amount of increase; y is the value of P/E ratio).

By 3), λ' on Muchincrease^- and Highyield^- can be established (table 5):

Table 5. λ on Muchincrease^- and Highyield^- .

<i>Index</i>	λ'
MUCHincrease^-	$\lambda'_1 = 0.236$
MUCHyield^-	$\lambda'_2 = 0.338$

3.5 Reasoning and Decision-Making of the Example

In above example, the decision rule is a kind of fuzzy reasoning. We use fuzzy production rule to determine the investment strategies in DR. The general form of fuzzy production rule [14] is as follows:

$$P_1, P_2, \dots P_m \rightarrow Q \mid \langle bd, (\tau_1, \tau_2, \dots \tau_m) \rangle$$

Where, P_i ($i = 1, 2, \dots, m$) are fuzzy sets that denotes premises, Q denotes conclusions or actions. bd ($0 \leq bd \leq 1$) is the belief degree of rule, τ_i ($0 \leq \tau_i \leq 1, i = 1, 2, \dots, m$) are threshold values of P_i . The fuzzy production rule means that “deduce Q from $P_1, P_2, \dots P_m$ with belief degree bd when $P_i(x)$ exceed τ_i for each i ”. Here, suppose that bd is 0.85. Next, we discuss an actual case on reasoning.

Due to DR and fuzzy production rule, DR can be represented:

- a) $\text{Muchincrease}^+(x) \wedge \text{Highyield}(x) \rightarrow \text{Stock}(\text{no}) \mid \langle 0.85, (0.764, 0.662) \rangle$
- b) $\text{Muchincrease}(x) \wedge \text{Highyield}^+(x) \rightarrow \text{Stock}(\text{yes}) \mid \langle 0.85, (0.764, 0.662) \rangle$
- c) $\text{Muchincrease}^-(x) \wedge \text{Highyield}^+(x) \rightarrow \text{Stock}(\text{more}) \mid \langle 0.85, (0.236, 0.662) \rangle$
- d) $(\text{Muchincrease}^+(x) \wedge \text{Highyield}^-(x)) \vee (\text{Muchincrease}(x) \wedge \text{Highyield}(x)) \rightarrow \text{Stock}(\text{little}) \mid \langle 0.85, (0.764, 0.662, 0.338) \rangle$
- e) $(\text{Muchincrease}^-(x) \wedge \text{Highyield}^-(x)) \vee (\text{Muchincrease}^+(x) \wedge \text{Highyield}^+(x)) \rightarrow \text{Stock}(\text{half}) \mid \langle 0.85, (0.236, 0.338, 0.764, 0.662) \rangle$

Take Kingdream as an example. Its amount of increase is 90.90%. Its P/E ratio is 82.89 times. According to (6) and (8), we can get: $\text{Muchincrease}(90.90) \approx 0.194 \leq 1-\lambda_1$. By 2) in 3.3, it indicates that Kingdream has small amount of increase. And $1-\lambda_2 \leq \text{Highyield}(82.89) = 0.402 \leq \lambda_2$. According to 3) in 3.3, it indicates that Kingdream has moderate P/E ratio. And, fuzzy production rule d) implies that investors can invest a small part of the capital in Kingdream.

Similarly, the decision results of all the stocks in the table 1 can be deduced. Because of space restrict, this paper gives part of the decision-making results merely (table 6).

From the table 6, we find that some stocks have the same decision result. In order to get more accurate result, we need to locate the most valuable stock from the same result. For this, we take weighted method. Suppose the weight of price change and P/E ratio has 0.5 weighting in stoke decision-making. Then, take decision-result ‘Stock(half)’ as an example.

Table 6.¹ Part of the Decision Results based on the Fuzzy Production Rules

Security (Abbreviation)	Result	Security (Abbreviation)	Result
HareonSolar	B	Kingdream	D
Sealand Securities	D	Science City	A
Zhejiang east	B	JOIN-IN (Holding)	E
VVFB	E	Jinfeng investment	E
ZHONG KE SAN HUAN	C	*ST Garden city	E
Ear Heng Yun A	C	SongDou shares	E
Rising Nonferrous	D	Tuopai Shede	E

TuopaiShede: $\text{Muchincrease}(91.16) \approx 0.196$, $\text{Highyield}^{\bar{d}}(39.65) = 0.829$, $f = 0.5 \times (0.196 + 0.829) = 0.5125$. VVFB: Since $\text{Muchincrease}(146.57) \approx 0.597$, $\text{Highyield}(102.18) \approx 0.505$, and the threshold value on $\text{Muchincrease}^{\sim}$ and Highyield^{\sim} are $\lambda'_1 = 0.236$, $\lambda'_2 = 0.338$ respectively. So, $\text{Muchincrease}^{\sim}(146.57)$ and $\text{Highyield}^{\sim}(102.18)$ are only satisfy (5). And $\text{Muchincrease}^{\sim}(146.57) = \text{Muchincrease}(146.57) \geq \lambda'_1$, $\text{Highyield}^{\sim}(102.18) = \text{Highyield}(102.18) \geq \lambda'_2$, $\text{Highyield}^{\bar{d}}(102.18) = 0.495$. So, $f = 0.546$. Thus, the following result can be calculated (table 7):

Table 7. Results computed through the weighted method

Security	f	Security	f
Tuopai Shede	0.5125	Suning Universal	0.4725
JOIN-IN(Holding)	0.5315	Guangzhou Pharmaceutical	0.386
SongDou shares	0.5525	MTC	0.48
Shanghai Construction	0.5398	Catic real estate	0.513
Jinfeng investment	0.496	XTC	0.445
Shandong Molong	0.4415	GOLD MANTIS	0.4235
JIUgui Liquor	0.4275	*ST Garden city	0.358
Chongqing industrial	0.4995	JIANFENG GROUP	0.5065
GuanCheng chase	0.515	VVFB	0.546

By the table 7, we know if investors decide to invest the half capital in stock, SongDou shares is the most valuable stock.

In the same way, if adopt the result Stock(yes), HareonSolar is a good choice. if adopt the result Stock(little), investors can choose Kingdream. If invest most of the capital in the stock, investors can choose ZHONG KE SAN HUAN

4 Conclusion

This paper has applied FSCOM in the stock investment decision-making. This paper researches real date of stocks. The instance of stock investment decision has practical

¹ A denotes Stock(no); B denotes Stock(yes); C denotes Stock(more); D denotes Stock(little); E denotes Stock(half).

significance. We improve their decision method when referring to a number of ways in paper [2] and [13]. This paper also adopts the method of weighting in process of decision-making and the decision process is more efficient. So, applying FSCOM in solving the objective actual problems with fuzziness and different negations is an effective way.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (6097315), the Fundamental Research Funds for the Central Universities (JUSRP51317B) and the Program for Innovative Research Team of Jiangnan University.

References

1. Pan, Z.: Three Kinds of Negation of Fuzzy Knowledge and Their Base of Set. *Chinese Journal of Computers* 35(7), 1421–1428 (2012) (in Chinese)
2. Pan, Z., Yang, L., Xu, J.: Fuzzy Set with Three Kinds of Negations and Its Applications in Fuzzy Decision Making. In: Deng, H., Miao, D., Lei, J., Wang, F.L. (eds.) *AICI 2011, Part I. LNCS (LNAI)*, vol. 7002, pp. 533–542. Springer, Heidelberg (2011)
3. Pan, Z.: Fuzzy Set with Three Kinds of Negations in Fuzzy Knowledge Processing. In: *Proceedings of 2010 International Conference on Machine Learning and Cybernetics (ICMLC 2010)*, vol. 5, pp. 2730–2735 (2010)
4. Zadeh, L.A.: Fuzzy Sets. *Information and Control* 8(3), 338–353 (1965)
5. Atanassov, K.: Intuitionistic Fuzzy Sets. *Fuzzy Sets and Systems* 20(1), 87–96 (1986)
6. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences* 11, 341–356 (1982)
7. Pan, Z., Zhu, W.: A New Cognition and Processing on Contradictory Knowledge. In: *Proceedings of IEEE-International Conference on Machine and Cybernetics*, vol. 3, pp. 1532–1537. IEEE Press, China (2006)
8. Yang, Y.: Weak, Also Has The Big Winner: 11 Stocks Have Doubled This Year. *Stock Market Trend Analysis Weekly* (24), 60–61 (2012) (in Chinese)
9. He, C.: Research on the Distribution Character of P/E Ratio in China's Stock Market and International Comparisons. *Economic Research Journal* (09), 74–81, 95 (2003) (in Chinese)
10. Wu, Y.: Application of the P/E Ratio in the Stock Investment Value Analysis. *Production Research* (21), 84–85 (2009) (in Chinese)
11. Zhang, L., Pan, Z.: Fuzzy Comprehensive Evaluation Based on Measure of Medium Truth Scale. In: *Proceedings of 2009 International Conference on Artificial Intelligence and Computational Intelligence (AICI 2009)*, vol. 2, pp. 83–87 (2009)
12. Li, J., Liao, H.: The Investment Tactics of the P/E Ratio: Appraise and Test. *Economic Management* (06), 73–79 (2007) (in Chinese)
13. Pan, Z., Wang, C., Zhang, L.: Three Kinds of Negations in Fuzzy Knowledge and Their Applications to Decision Making in Financial Investment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) *ICCCI 2010, Part II. LNCS (LNAI)*, vol. 6422, pp. 391–401. Springer, Heidelberg (2010)
14. Dung, P.M., Mancarella, P.: Production Systems Need Negation as Failure. *IEEE Transactions on Knowledge and Computer and Information Sciences* 11, 341–356 (1982)

Two-Phase Image Segmentation with the Competitive Learning Based Chan-Vese (CLCV) Model

Yanqiao Zhu, Anhui Wang, and Jinwen Ma^{*}

Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China
zhuyanqiao0413@gmail.com,
jwma@math.pku.edu.cn

Abstract. In this paper, we propose a competitive learning based Chan-Vese model (CLCV) for two-phase image segmentation by coupling the Chan-Vese model and the rival penalized competitive learning mechanism from the point of view of the cost function for the DSRPCL algorithm. Specifically, the CLCV model based approach to image segmentation incorporates the mechanism of rival penalized competitive learning into the evolution of the level set function so that there emerge certain repulsive forces between the foreground and background classes, which lead to more accurate segmentations of the image. Experimental results on several real-world images have validated the advantages of the proposed CLCV model over the original Chan-Vese model on integral segmentation, smooth boundaries and robustness to noises.

Keywords: Image Segmentation, Level Set, Competitive Learning, Chan-Vese Model.

1 Introduction

Image segmentation is one of the most important problems in image processing and computer vision, which refers to partitioning a digital image into some disjoint regions with homogeneous characteristics. Its aim is to simplify and change the representation of an image into something more meaningful and easier to analyze.

Regarding automatic image segmentation, i.e., image segmentation without any prior information, there have been a variety of theories and methods. Edge-based methods utilize edge detectors such as the Sobel and Laplace operators to locate edges in an image and then obtain the object's boundaries [1, 2]. Further efforts are usually paid to discover true boundaries and link broken ones. Region-based algorithms, for instance, region growing methods, pursue interested object regions directly [3, 4], but the proper initialization of seeds and possible over-segmentation are great challenges. Unsupervised clustering algorithms like K-means and competitive learning can also

^{*} Corresponding author.

be employed for image segmentation [5, 6]. However, integral segmentation is hard to obtain since they operate on the single pixel level and neglect the spatial continuity of objects.

Deformable models are also frequently adopted in image segmentation. Starting with an initial curve C , the snake model can gradually deform C to the boundary of the object. The evolution of C tries to minimize the overall energy composed of internal energy related to curve smoothness and external energy relied on image gradients [7]. Due to the parameterized representation of the curve, classical active contour methods have difficulties in dealing with topology changes of the curve, for instance, merging and splitting. The level set method [8] is then adopted to solve these challenges by formulating the curve C implicitly as the zero level set of a Lipschitz function $\phi(x, y)$. The evolution of C is then transformed to the evolution of $\phi(x, y)$ according to the associated Euler-Lagrange equation. The geodesic active model [9] provides such an example. For these models, the curve stops evolving when edges are discovered. Therefore they can find only objects with clear edges defined by image gradients.

Based on the Mumford-Shah segmentation model [10], Chan and Vese proposed a model which did not rely on image gradients to stop the level set evolution so that it could detect contours with or without edges [11, 12]. This method greatly extends the processing ability of deformable models and thus is widely applied. However, Chan-Vese model based approaches still have great difficulties in processing complex images. First, results can be messy due to the existence of noise and great morphology variation. Second, local segmentations are common which fail to delineate the whole integral object.

To overcome these difficulties, we try to incorporate the competitive learning scheme into the Chan-Vese model. Actually, competitive learning methods are famous for their simplicity and power, and thus widely applied in various practical problems. Especially, the rival penalized competitive learning (RPCL) algorithm [13] introduced a de-learning procedure or certain repulsive forces into the competitive learning paradigm such that it had the ability of automatically determining the true number of clusters for a dataset [13]. Ma et al. later established a theoretical framework for the heuristic RPCL algorithm using a cost function approach [14, 15]. Other than the ability of automatic cluster number determination, the de-learning process in these methods can also help to provide better inter-class boundaries, which is less mentioned but crucial for the segmentation of complex real world images. Therefore, we incorporate the mechanism of rival penalized competitive learning into the traditional Chan-Vese model and propose the competitive learning based Chan-Vese (CLCV) model. In fact, the de-learning procedure and repulsive forces are integrated in the level set evolution to provide more accurate segmentations.

The rest of the paper is organized as follows. Section 2 proposes the new competitive learning based Chan-Vese (CLCV) model. Section 3 presents experimental results on nine real-world images. Finally, brief conclusions and discussions are provided in Section 4.

2 Methodology

2.1 Preliminary Knowledge

In order to propose the competitive learning based Chan-Vese model, we begin to briefly introduce the Chan-Vese model and the Distance Sensitive RPCL (DSRPCL) algorithm as well as its cost function.

The Chan-Vese Model. The Chan-Vese model tries to minimize the following cost function:

$$\begin{aligned}
 F_1(c_1, c_2, C) = & \mu \cdot \text{Length}(C) + v \cdot \text{Area}(\text{inside}(C)) \\
 & + \lambda_1 \iint_{\text{inside}(C)} \|u_0(x, y) - c_1\|^2 dx dy \\
 & + \lambda_2 \iint_{\text{outside}(C)} \|u_0(x, y) - c_2\|^2 dx dy.
 \end{aligned} \tag{1}$$

Here C is the curve representing the dividing boundary between the foreground and the background, c_1 and c_2 are two constants which try to approximate the image function $u_0(x, y)$ inside and outside C , respectively, while $\lambda_1 \geq 0, \lambda_2 \geq 0, \mu \geq 0, v \geq 0$ are regularization parameters. In most cases, λ_1 and λ_2 are set to 1 and v is set to 0. The item $\text{Length}(C)$ imposes the smoothness requirement on the curve, while the two integration terms impose the data fitting demand.

When $\lambda_1 = \lambda_2 = \lambda, v = 0$, the Chan-Vese model degenerate to the two-phase minimum partition problem, which is a simplified version of the Mumford-Shah model [10] with only piecewise level functions considered.

The DSRPCL Algorithm. The DSRPCL algorithm introduces the rival penalized competitive learning mechanism into the ordinary K-means scheme [14, 15] by modifying the cost function as:

$$F_2(W) = \frac{1}{2} \sum_{k=1}^K \sum_{f_s \in G_k} \|f_s - w_k\|^2 + \frac{2}{P} \sum_{k=1}^K \sum_{f_t \notin G_k} \|f_t - w_k\|^{-P}, \tag{2}$$

where G_k is the set of samples in the k -th cluster, f_s is the s -th sample, $W = [w_1, w_2, \dots, w_K]$, w_k is the weight vector of unit k , $k = 1, 2, \dots, K$. S represents the number of samples and K is the number of units. The parameter $P > 0$ is used to control the strength of penalty to the rivals. In fact, the second term of this function provides certain repulsive forces during the learning process, which corresponds to the de-learning process in the RPCL algorithm. This work has established a theoretical framework for its heuristic counterpart.

Other than the commonly referred ability of automatically determining the number of clusters, the mechanism of de-learning represented by the second term in Eq. (2)

can also help to find better inter-class boundaries, especially in the case of overlapping [14, 15]. Therefore, it can play an important role in the segmentation of objects with large internal morphology variation and smooth transition to the background in noisy images.

2.2 The Competitive Learning Based Chan-Vese Model

We now propose the competitive learning based Chan-Vese (CLCV) model which tries to couple the Chan-Vese model and the DSRPCL mechanism such that the integration of competitive learning in the curve evolution will help to build more accurate objects boundaries and provide better segmentation results. Mathematically, it can be formulated to minimize the following modified cost function:

$$\begin{aligned}
 F(c_1, c_2, C) = & \mu \cdot \text{Length}(C) + v \cdot \text{Area}(\text{inside}(C)) \\
 & + \lambda_1 \iint_{\text{inside}(C)} \left(\frac{1}{2} \|u_0(x, y) - c_1\|^2 + \frac{2}{P} \|u_0(x, y) - c_2\|^{-P} \right) dx dy \quad (3) \\
 & + \lambda_2 \iint_{\text{outside}(C)} \left(\frac{1}{2} \|u_0(x, y) - c_2\|^2 + \frac{2}{P} \|u_0(x, y) - c_1\|^{-P} \right) dx dy.
 \end{aligned}$$

Here the parameters $\lambda_1 \geq 0, \lambda_2 \geq 0, \mu \geq 0, v \geq 0$ and $P > 0$ are inherited from the Chan-Vese model and the DSRPCL model, respectively.

Clearly, the proposed CLCV model differs from the original Chan-Vese model at the data fitting parts. The items $\frac{2}{P} \|u_0(x, y) - c_1\|^{-P}$ and $\frac{2}{P} \|u_0(x, y) - c_2\|^{-P}$ are introduced here to incorporate the competitive learning scheme, as in the DSRPCL algorithm. This actually imposes repulsive forces between the foreground and background classes. Smaller P implements stronger interactions between them. Although these interactions may cause inaccurate approximation of the image function, they will at the same time provide better separation between different classes, especially in complex cases with overlapping or unclear boundaries. Thus, it holds great potential in providing better segmentation results than the original Chan-Vese model.

Formulating the proposed CLCV model with the level set method, we have:

$$\begin{aligned}
 F(c_1, c_2, \phi) = & \mu \iint_{\Omega} \|\nabla H(\phi)\| dx dy + v \iint_{\Omega} H(\phi) dx dy \\
 & + \lambda_1 \iint_{\Omega} \left(\frac{1}{2} \|u_0(x, y) - c_1\|^2 + \frac{2}{P} \|u_0(x, y) - c_2\|^{-P} \right) H(\phi) dx dy \quad (4) \\
 & + \lambda_2 \iint_{\Omega} \left(\frac{1}{2} \|u_0(x, y) - c_2\|^2 + \frac{2}{P} \|u_0(x, y) - c_1\|^{-P} \right) (1 - H(\phi)) dx dy,
 \end{aligned}$$

where Ω is the image domain, $H(z)$ is the Heaviside function that takes 1 when $z \geq 0$ and takes 0 otherwise, $\phi = \phi(x, y)$ is the level set function. One common choice for ϕ is the signed distance function. The curve C in Eq. (3) is now represented by the

zero level set of ϕ , i.e., $C = \{(x, y) : \phi(x, y) = 0\}$. To derive the evolution scheme of ϕ by solving the associated Euler-Lagrange equation [11, 12], a regularized version of H is adopted:

$$H_\varepsilon(z) = \begin{cases} 1, & \text{if } z > \varepsilon; \\ 0, & \text{if } z < -\varepsilon; \\ \frac{1}{2} \left(1 + \frac{z}{\varepsilon} + \frac{1}{\pi} \sin\left(\frac{\pi z}{\varepsilon}\right) \right), & \text{if } |z| \leq \varepsilon, \end{cases} \quad (5)$$

where $\varepsilon > 0$ is a small number. The corresponding regularized model is then:

$$\begin{aligned} F_\varepsilon(c_1, c_2, \phi) &= \mu \iint_\Omega \|\nabla H_\varepsilon(\phi)\| dx dy + v \iint_\Omega H_\varepsilon(\phi) dx dy \\ &+ \lambda_1 \iint_\Omega \left(\frac{1}{2} \|u_0(x, y) - c_1\|^2 + \frac{2}{P} \|u_0(x, y) - c_2\|^{-P} \right) H_\varepsilon(\phi) dx dy \\ &+ \lambda_2 \iint_\Omega \left(\frac{1}{2} \|u_0(x, y) - c_2\|^2 + \frac{2}{P} \|u_0(x, y) - c_1\|^{-P} \right) (1 - H_\varepsilon(\phi)) dx dy. \end{aligned} \quad (6)$$

The Euler-Lagrange equation for ϕ is obtained by introducing a time parameter $t \geq 0$ and minimize $F_\varepsilon(c_1, c_2, \phi)$ with respect to ϕ when keeping c_1 and c_2 fixed. Specifically,

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \delta_\varepsilon(\phi) \left\{ \mu \nabla \left(\frac{\nabla \phi}{\|\nabla \phi\|} \right) - v - \lambda_1 \left(\frac{1}{2} \|u_0 - c_1\|^2 + \frac{2}{P} \|u_0 - c_2\|^{-P} \right) \right. \\ &+ \lambda_2 \left. \left(\frac{1}{2} \|u_0 - c_2\|^2 + \frac{2}{P} \|u_0 - c_1\|^{-P} \right) \right\} = 0 \text{ in } (0, \infty) \times \Omega, \\ \phi(0, x, y) &= \phi_0(x, y) \text{ in } \Omega, \\ \frac{\delta_\varepsilon(\phi)}{\|\nabla \phi\|} \frac{\partial \phi}{\partial \mathbf{n}} &= 0 \text{ on } \partial \Omega, \end{aligned} \quad (7)$$

where $\delta_\varepsilon(z) = H_\varepsilon'(z)$, $\phi_0(x, y)$ is the initial level set, and \mathbf{n} denotes the exterior normal to the image boundary. Note that the competitive learning scheme and repulsive forces are embodied in the evolution of ϕ . The equation can then be discretized and solved using the gradient descent method.

As for the update of c_1 and c_2 , extra efforts are required to solve the corresponding optimization problem when ϕ is fixed at time t . However, since c_1 and c_2 are originally designed to approximate the original image function u_0 and the competing mechanism is already incorporated into the level set evolution, one reasonable alternative to

update c_1 and c_2 is to use the mean features in corresponding regions as in the Chan-Vese model. For more details, readers can refer to [11, 12].

In summary, the image segmentation learning algorithm based on the proposed CLCV model includes four main steps:

1. Initialize $\phi(0, x, y)$ by $\phi_0(x, y)$, and set $t = 0$.
2. Update c_1^t and c_2^t using the mean features in corresponding regions.
3. Compute ϕ^{t+1} according to Eq. (7) and reinitialize it to the signed distance function.
4. If ϕ is stable or the maximum iteration number is reached, stop; otherwise, set $t = t + 1$ and repeat steps 2-4.

3 Experimental Results

3.1 Feature Extraction

For each pixel represented by its location (i, j) in a given image u_0 , a feature vector composed of eight features is extracted, denoted by $X_{ij} = [X_{ij}(1), X_{ij}(2), \dots, X_{ij}(8)]$.

$[X_{ij}(1), X_{ij}(2), X_{ij}(3)]$ represent color space features. Here, the most discriminative 3 channels in term of entropy are selected out of the mixed color spaces, i.e., [R, G, B, L, U, V, Y, Cb, Cr], as the color space features. It should be noted that [R, G, B], [L, U, V], [Y, Cb, Cr] are the RGB, CIE 1976 (L^* , u^* , v^*) and YCbCr representation of images, respectively. $[X_{ij}(4), X_{ij}(5), X_{ij}(6)]$ are the texture features computed with the method described in [16] which utilizes wavelet transform at multiple scales. $[X_{ij}(7), X_{ij}(8)]$ are the regularized coordinates used as spatial features.

The justification for introducing spatial features lies in the fact that objects of interest are usually continuous regions, or composed of several continuous regions. Therefore, spatial features are adopted here to impose the spatial continuity constraint. The normalized feature vectors are then adopted in the following experiments.

3.2 Results and Analysis

We validated the proposed two-phase CLCV approach on several real-world images from the Berkeley segmentation database [17]. For comparison, we also tested the Chan-Vese model on the same images with the original RGB features and the selected features. The key parameters were set as follows in the experiments: $\lambda_1 = 1, \lambda_2 = 1, \mu = 0.3, v = 0, P = 0.55$. A rectangular with 1/3 of the size of a given image was placed in the image center and its border was set as the initial zero level set.

As shown in Fig. 1, segmentation results for the Chan-Vese model with original RGB features were not satisfying. The segmented objects were not complete and

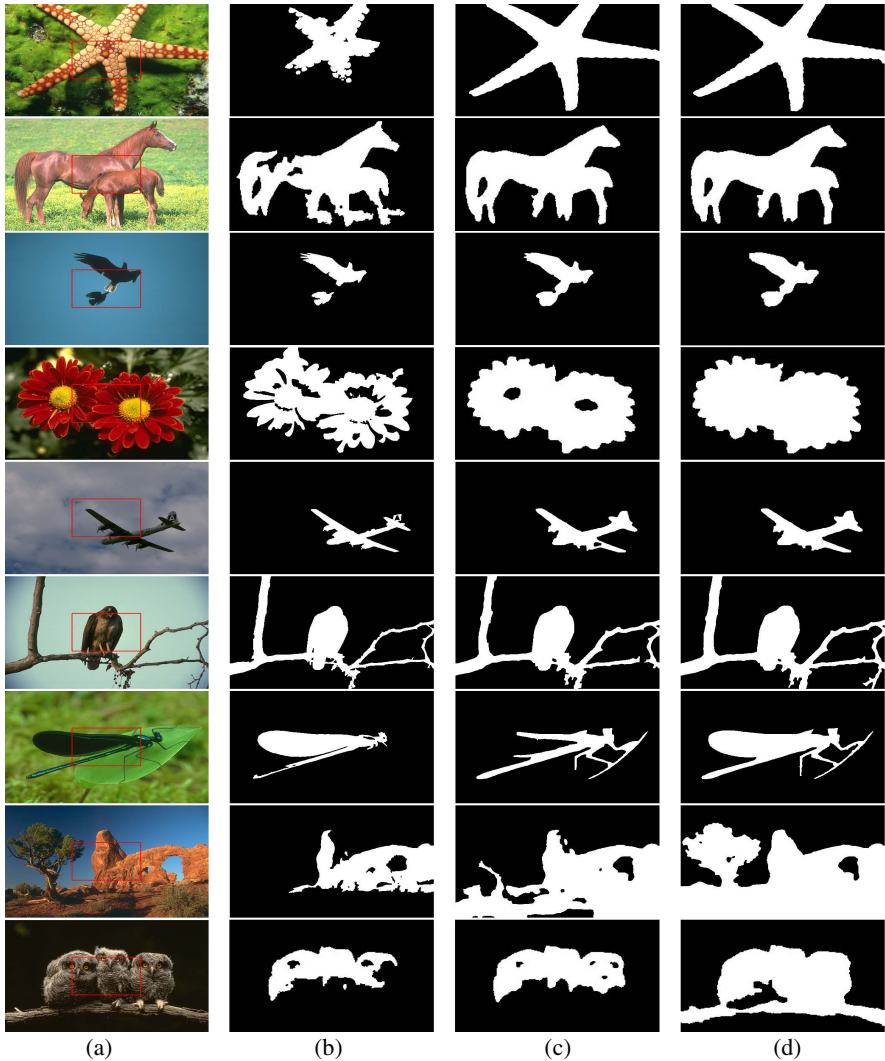


Fig. 1. Experimental results on nine real-world images. (a) Original images, red lines indicated initial contours, (b) segmentation results of the Chan-Vese model using RGB features, (c) segmentation results of the Chan-Vese model using the selected features, (d) segmentation results of the CLCV model using the selected features.

integral. For example, certain parts of the starfish, the horses, the flowers and the plane were missing. The background was also misclassified as objects in the horses and flowers images. These were mainly caused by the complex morphologies of image scenes. Intuitively, shadow regions in the horse image looked more like parts of horse rather than grass and vice versa. Although not too much noise was presented in these images, local segmentations were still presented when only RGB color space features were considered.

The Chan-Vese method with our selected features performed much better. The starfish and horses were completely recovered. However, there could still be problems with certain sensitive regions of objects and caused holes in the flowers image and the plane image. The cases were even worse for the landscape and owls images. The proposed CLCV model, however, greatly overcame these problems and successfully reconstructed the whole objects. The segmentation results were integral, and detected boundaries were more smooth and accurate.

For the quantitative evaluation purpose, we calculated the probabilistic rand index (PRI) [18] values on segmentation results of each method. The average PRI values for the Chan-Vese model with RGB features, Chan-Vese model with selected features and CLCV model with selected features were 0.5428, 0.5756 and 0.5997, respectively, which manifested the advantages of the selected features and the proposed CLCV model. It was clear that all the three PRI values were not very high, which might be caused by the fact the PRI favored over-segmentation while we were performing two-phase segmentation. In fact, the PRI values of all three cases were over 0.9 for the hawk image in which manual segmentations were more simple and consistent. Therefore, we might need to search for the other ways to better measure the integrity of segmentation results which was our focus of the study.

By comparing Eq. (1) and Eq. (3), one can see that the difference between the original Chan-Vese model and the proposed CLCV model lies in the competitive scheme introduced in the CLCV model. Experimental results on nine real-world images verified its ability in providing better separations for different kinds of images.

4 Conclusions and Discussions

In this paper we have proposed a novel model which incorporates a rival penalized competitive mechanism into the traditional Chan-Vese model. Experimental results on several real-world images validated the effectiveness of the proposed competitive learning based Chan-Vese model. Accurate boundaries were delineated and integral segmentation results were obtained, thanks to the de-learning and repulsive forces introduced in the level set evolution. This model considerably overcomes the difficulties which the Chan-Vese model has been struggling with when processing real-world images with complex morphologies.

In the future, we will further improve the performance of the model and investigate the proper settings of the parameters. Currently the segmentation result of the proposed CLCV method is usually a little bit larger than the object due to the adopted texture features. More meaningful features will be tested to provide corrections for segmentation. Also, we would like to generalize the model into multi-phase image segmentation. As for the multi-phase case, automatic determination of objects number in a given image will also be explored.

Acknowledgment. This work was supported by the Natural Science Foundation of China for grant 61171138.

References

1. Gonzalez, R., Woods, R., Eddins, S.: *Digital Image Processing Using MATLAB*. Prentice Hall, Upper Saddle River (2004)
2. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an Image Edge Detection Filter Using the Sobel Operator. *IEEE Journal of Solid-State Circuits* 23(2), 358–367 (1988)
3. Adams, R., Bischof, L.: Seeded Region Growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(6), 641–647 (1994)
4. Mancas, M., Gosselin, B., Macq, B.: Segmentation Using a Region-Growing Thresholding. In: *Proceedings of SPIE 5672, Image Processing: Algorithms and Systems IV* (2005)
5. Hartigan, J.: *Clustering algorithms*. John Wiley & Sons, Inc., New York (1975)
6. Boujemaa, N.: Generalized Competitive Clustering for Image Segmentation. In: *19th International Conference of the North American Fuzzy Information Processing Society*, pp. 133–137 (2000)
7. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. *International Journal of Computer Vision* 1(4), 321–331 (1988)
8. Osher, S., Sethian, J.A.: Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *Journal of Computational Physics* 79(1), 12–49 (1988)
9. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. *International Journal of Computer Vision* 22(1), 61–79 (1997)
10. Mumford, D., Shah, J.: Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems. *Communications on Pure and Applied Mathematics* 42(5), 577–685 (1989)
11. Chan, T., Vese, L.: An Active Contour Model without Edges. In: Nielsen, M., Johansen, P., Fogh Olsen, O., Weickert, J. (eds.) *Scale-Space 1999*. LNCS, vol. 1682, pp. 141–151. Springer, Heidelberg (1999)
12. Chan, T.F., Sandberg, B.Y., Vese, L.A.: Active Contours without Edges for Vector-Valued Images. *Journal of Visual Communication and Image Representation* 11(2), 130–141 (2000)
13. 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)
14. Ma, J., Wang, T., Xu, L.: Convergence Analysis of Rival Penalized Competitive Learning (RPCL) Algorithm. In: *2002 International Joint Conference on Neural Networks*, vol. 2, pp. 1596–1601 (2002)
15. Ma, J., Wang, T.: A Cost-Function Approach to Rival Penalized Competitive Learning (RPCL). *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 36(4), 722–737 (2006)
16. Unser, M.: Texture Classification and Segmentation Using Wavelet Frames. *IEEE Transactions on Image Processing* 4(11), 1549–1560 (2002)
17. 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: *8th IEEE International Conference on Computer Vision*, vol. 2, pp. 416–423 (2001)
18. Unnikrishnan, R., Hebert, M.: Measures of Similarity. In: *7th IEEE Workshops on Applications of Computer Vision*, vol. 1, pp. 394–400 (2005)

Adaptive Step Searching for Solving Stochastic Point Location Problem

Tongtong Tao, Hao Ge, Guixian Cai, and Shenghong Li*

Department of Electronic Engineering, Shanghai Jiao Tong University,
800 Dongchuan Road, Min Hang, Shanghai 200240, P.R. China
shli@sjtu.edu.cn

Abstract. A novel algorithm named Adaptive Step Searching (ASS) is presented in the paper to solve the stochastic point location (SPL) problem. In the conventional method [1] for the SPL problem, the tradeoff between the convergence speed and accuracy is the main issue since the searching step of learning machine (LM) in the method is invariable during the entire searching. In that case, in ASS, LM adapts the step size to different situations during the searching. The convergence speed has been improved significantly with the same accuracy comparing to previous algorithms.

Keywords: Stochastic Point Location, Learning Automata.

1 Introduction

Point location problem considers a robot (learning machine (LM), algorithm, etc.) searching for a point or a parameter in an interval. To assist the LM to search for the point, we assume that the LM can communicate with the environment, and that the environment will provide the feedback information about which direction to go. Then, the LM will move based on the feedback information.

If the environment is a deterministic environment, that is to say, the information provided by the environment is deterministic, and this problem can be classified as deterministic point location (DPL) problem. Correspondingly, a stochastic environment (SE) means that wrong information is possible, for instance, when LM should go to the right, the SE may suggest it to go to the left with a nonzero probability, vice versa. This problem can be classified as stochastic point location (SPL) problem. Unlike the DPL problem, the error possibility of the feedback from the SE in the SPL problem would make this problem more challenging and fascinating. In the SPL problem pioneered by Oommen[1,2,3,4], the LM attempts to find a point on the line with stochastic (possibly wrong) feedback from the environment. The main shortcoming of the prior state of art solution reported in [1] is that the steps are always tough. The tradeoff between convergence speed and accuracy is the main issue since the searching step in the method is invariable during the entire searching. Choosing a small step

* Corresponding author.

size will lead to higher accuracy, but slower speed of convergence. However, increasing the step size will improve the convergence speed, meanwhile reduce the accuracy.

Furthermore, the stochastic environment (SE) can be classified into two types [2][3]. We assume that the SE suggests the correct information with the probability p , and error information with probability $1-p$. The first type of SE is called an informative environment, in which the probability of the SE suggesting correct information is bigger than 0.5 ($p > 0.5$); the second type is called a deceptive environment, in which the probability of the SE suggesting the correct information is smaller than 0.5 ($p < 0.5$).

In this paper, the SPL problem in an informative environment is concerned and we introduce an algorithm named Adaptive Step Searching (ASS) to solve this problem. In the ASS, we extended the algorithm in [1] by discretizing the interval into adaptive steps. The heart of the strategy is how to design appropriate step size to search for the optimal point in order to enhance the convergence speed.

Apart from the importance of the SPL problem itself, LA has a myriad of applications in solving optimization problems, such as network and communications [5], traffic control[6], graph partition[7] and so on.

2 Related Work

Oommen proposed the algorithm for solving this problem in an informative Environment ($p > 0.5$) by discretizing the search space [1]. Without loss of generality, we can regard the search space as a unit interval. First, divide the search space into N steps $\{0, 1/N, 2/N, \dots, (N-1)/N, 1\}$, and a bigger N will lead to a more precise convergence to the unknown optimal λ^* . Whenever the SE suggests the LM to go to the left (right), it moves to the left (right) by a single step ($1/N$) in the space. The environment is stochastic, so when the unknown λ^* is on the right side of LM, the SE may probably suggest the LM to go to the left wrongly. However, Oommen has analyzed the algorithm, and concluded that if the environment is informative, the LM can converge to the parameter λ^* , and the learning can always be asymptotically optimal.

As mentioned above, the main drawback of the solution reported in [1] is the tradeoff between convergence speed and accuracy. Increasing the step size will improve the converging speed, meanwhile the accuracy is correspondingly decreased. However, only a large N can lead to a precise convergence.

To deal with the tradeoff mentioned above, Oommen has proposed Hierarchical Stochastic Searching on the Line (HSSL) in [4], which structures the space as a binary tree. The search is arranged in the binary tree from one node to another with depth $D = \log_2(N)$, and the resolution becomes more accurate at higher levels. Define $\Delta = [\sigma, \lambda)$ as the current search interval containing λ^* . At the beginning, $\Delta = \Delta\{0, 1\} = [0, 1)$, and parent $\Delta\{d, j\}$ (d indicates the depth of the node in the binary tree and j indicates the index of node among the d -depth nodes) can be divided into two equal intervals, left child $\Delta\{d+1, 2j-1\}$ and right child $\Delta\{d+1, 2j\}$. Choose three sampled points, the left point, midpoint and right point of the current interval, and decide next search interval based on the suggestion from the SE for the three sampled points.

The HSSL has a better performance than the conventional SPL with three LMs relative to the conventional SPL. In this paper, we present Adaptive Step Search (ASS) algorithm to solve the SPL problem, and verify it can improve the convergence speed comparing to the conventional SPL under different conditions. In the slowly varying non-stationary environment, it is superior to the HSSL greatly.

3 Adaptive Step Searching (ASS)

The algorithm we propose rooted from the conventional SPL in [1] by discretizing the interval and the pertinent contribution is that the LM in our scheme searches the interval for the optimal λ^* using the adaptive step size.

Let N_{min} as the smallest resolution, $1/N_{min}$ as the biggest search step size, N_{max} as the biggest resolution, and $1/N_{max}$ be the smallest search step size. $N_{current}$ is defined as the LM's current resolution which is between N_{min} and N_{max} , and when the optimal λ^* is far away from the current point, larger step size ($1/N_{current}$) is chosen for LM to approach the optimal point faster using smaller $N_{current}$ ($N_{min} < N_{current} < N_{max}$). On the other hand, when optimal λ^* is nearby, higher resolution $N_{current}$ ($N_{min} < N_{current} < N_{max}$) is used to reach the optimal point more accurately.

The heart of ASS algorithm is to adapt the current resolution $N_{current}$ to the environment during the searching. Different from the algorithms in [1] [4], the LM in ASS needs 2bits memory to record the SE feedback information of the last two moments. The memory can help to decide the current resolution. Digit "0" represents that the SE suggests the LM to move toward left and digit "1" represents right. The state in the memory can be {00, 01, 11, 10}, and the state transition is illustrated as Fig.1, where the dotted line represents the left (0) suggestion and real line represents the right (1) suggestion. The two bits in the memory and the current feedback information compose the decision conditions {000, 001, 010, 011, 100, 101, 110, and 111}. Decision table for choosing current resolution is presented in Table 1 to illustrate ASS algorithm clearly. In Table 1, $\lambda(n-1)$ denotes the value of the moment $n-1$, and the step size for this moment can be taken according to the different conditions.

The items in the decision table can be classified into three cases. At the beginning, the state in the memory will randomly set to {00, 01, 11, 10}. In the first case, if at time $n-2$, the SE suggests the LM to go the left (right), the memory will record the feedback information 0 (1). At time $n-1$, the SE suggests the LM to move to the left (right) again. LM moves to the left (right) with step $1/N_{n-1}$, then 0 (1) is also written into the memory. The memory in LM presents 00 (11). Then, at the time n , the SE suggests the LM to move to the left (right) once again. While at this time, LM moves to the left (right) with a step $1/N_n$ and the memory will record the feedback information 0 (1), the knowledge in the memory is still 00 (11). In this situation where the SE always suggests the LM to move to the left (right), it is assumed that the optimal λ^* is far away. In that case, step size should be increased to approach the optimal location more quickly. For ASS, when the condition is 000 or 111, resolution value is reduced by half, that is to say, $N_n = N_{n-1}/2$.

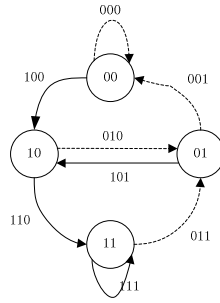


Fig. 1. State transition

Table 1. Decision table to choose the current step size

Condition	Current Resolution	Current Value
000	$N_n = N_{n-1}/2$	$\lambda(n) = \lambda(n-1) - 1/N_n$
001	$N_n = N_{n-1}$	$\lambda(n) = \lambda(n-1) - 1/N_n$
010	$N_n = N_{n-1} * 2$	$\lambda(n) = \lambda(n-1) - 1/N_n$
011	$N_n = N_{n-1}$	$\lambda(n) = \lambda(n-1) - 1/N_n$
100	$N_n = N_{n-1}$	$\lambda(n) = \lambda(n-1) + 1/N_n$
101	$N_n = N_{n-1} * 2$	$\lambda(n) = \lambda(n-1) + 1/N_n$
110	$N_n = N_{n-1}$	$\lambda(n) = \lambda(n-1) + 1/N_n$
111	$N_n = N_{n-1}/2$	$\lambda(n) = \lambda(n-1) + 1/N_n$

In the second case, the SE suggests the LM to go the left (right), next time it changes the suggestion, and then it changes the suggestion once again. In this case, it is supposed that the λ^* is nearby and the step size is too big. Thus the current resolution should be increased to reach the optimal location more accurately. For ASS, when the condition is 010 or 101, current resolution N_n is enlarged by half to $N_{n-1} * 2$.

In all the other cases, the conditions are 001, 011, 100 and 100, respectively. For these situations, the feedback information from SE is unstable so that whether the optimal λ^* is nearby or far away from current location cannot be determined. To be conservative, we assume some errors have happened in the SE suggestion, it is fair to make N_n equal to N_{n-1} .

4 Simulations and Results Analysis

Two sets of experiments are presented to illustrate the convergence property of the ASS clearly, although multiple experiments have been carried out.

The first set of experiment is conducted to demonstrate ASS’s correct convergence under different environments (diverse values of p and N_{max}). In the second set of experiment, we compare the performance of ASS with the previous SPL solutions.

In the first set of experiment, we record the value of $E[\lambda(\infty)]$ for diverse p and N_{max} , and the results are presented in Table 2. The $E[\lambda(\infty)]$ were evaluated by performing 5000 experiments for each pair.

The convergence of ASS method can be demonstrated by this set of experiment. Table 2 shows even when p is as low as 0.7, the algorithm can still converge to the optimal, and results gets better if N_{max} is increased. When $p=0.8$, $N_{max} =1024$, the estimate of ASS converges to the optimal λ^* with only 0.013% error.

Table 2. The values of $E[\lambda(\infty)]$ for diverse p and N_{max} when the optimal λ^* is 0.9123.

$\text{Log}_2(N_{max})$	P=0.7	P=0.8	P=0.9
3	0.86365	0.904575	0.926
4	0.8726625	0.8954875	0.904725
5	0.8955375	0.9156	0.92070625
6	0.89925625	0.91178125	0.913740625
7	0.89995625	0.9091296875	0.9097265625
8	0.90440703125	0.9116890625	0.9124671875
9	0.908187890625	0.912908984375	0.913234375
10	0.9086021484375	0.9124978515625	0.9125373046875
11	0.91016826171875	0.9123345703125	0.9123341796875
12	0.91002734375	0.912111425781250	0.912229541015625

In the second set of experiment, the p was fixed in 0.8, and ASS was compared with the previous SPL algorithms. The results are shown in Fig.2 and Fig.3. For each simulation, 1000 repeated experiments have been performed.

As shown in Fig.2 the ASS outperforms the conventional SPL solution. In Fig.2 (a), the N_{max} of ASS is set to 256 and the resolution of the conventional SPL [1] (i.e. N) is also set to 256, while the optimal λ^* alternates between 0.9123 and 1-0.9123 every 400 iterations. In Fig.2 (b), the N_{max} of ASS is set to 1024 and the resolution of the conventional SPL (i.e. N) is also set to 1024, while the optimal λ^* alternates between 0.9123 and 1-0.9123 every 1500 iterations.

From Fig.2 (a), in the first 400 iterations, the ASS only needs 30 iterations to reach 99% of λ^* , but the conventional SPL solution needs more than 180 iterations. In the second 400 iterations, when the optimal λ^* changes to 0.0877 (1-0.9123). The ASS reaches 95% of λ^* within 60 iterations while the conventional SPL solution needs 380 iterations.

The similar results can be obtained from Fig.2 (b) whose resolution is 1024. In the first 1500 iterations, the ASS only needs 27 iterations to reach 99% of λ^* , however,

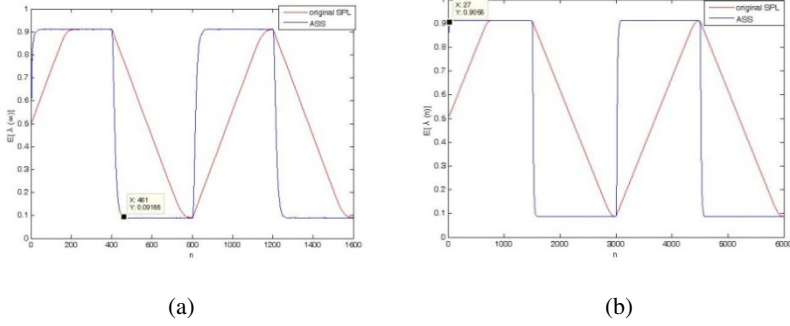


Fig. 2. The learning property of the ASS and the conventional SPL algorithm with different resolution: (a) $N_{max} = N = 256$, and (b) $N_{max} = N = 1024$.

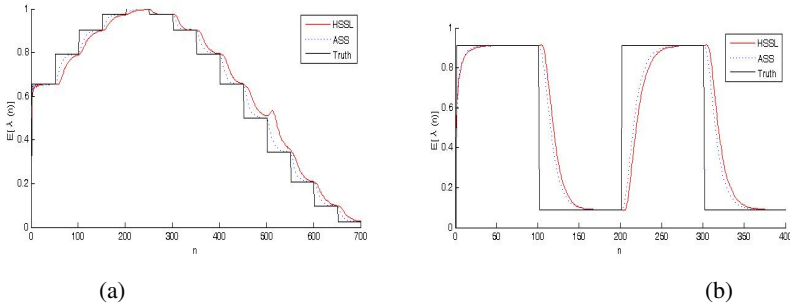


Fig. 3. The learning characteristics of the ASS and the HSSL algorithm in non-stationary environment: (a) slowly varying, $p=0.8$, $N=1024$, and (b) fiercely varying, $p=0.8$, $N=256$

the conventional SPL solution needs 700 iterations to reach 99% of λ^* . In the second 1500 iterations, when the optimal λ^* changes to 0.0877. The ASS can reach 95% of λ^* within 77 iterations while the conventional SPL solution needs 1440 iterations. Due to the increase of N , the convergence speed of the conventional SPL solution decreases seriously, however, the increase of N has little influence to the ASS.

As shown in Fig.3, the ASS was compared with the HSSL [4] in different type of non-stationary environment. For the slowly varying non-stationary environment, the ASS is superior to the HSSL greatly, which is shown in Fig.3 (a). For the fiercely varying non-stationary environment, the performance of the HSSL and ASS are almost the same, which is shown in Fig.3 (b). A two bits memory is needed in ASS, but the HSSL needs three LMs to judge each step. We can say the compare is quite fair.

5 Conclusion

The ASS introduced in the paper shows great advantages for solving the SPL problem. By adapting the step size in different situations, the LM in the ASS can converge

to the optimal value fast and accurately. Experiments show that the ASS performs better than the previous SPL solutions under different conditions. As the advantages shown above, the ASS can be widely used in solving optimization problems, such as network and traffic control and so on.

Acknowledgment. The authors would like to acknowledge the support from National Natural Science Foundation of China (61071152, 61271316), 973 Program (2010CB731403, 2010CB731406, 2013CB329605) of China and Chinese National "Twelfth Five-Year" Plan for Science & Technology Support (2012BAH38B04), Key Laboratory for Shanghai Integrated Information Security Management Technology Research, and Chinese National Engineering Laboratory for Information Content Analysis Technology.j

References

1. Oommen, B.J.: Stochastic Searching on the Line and Its Applications to Parameter Learning in Nonlinear Optimization. *IEEE Transaction Systems, Man, and Cybernetics* 27(4), 733–739 (1997)
2. Oommen, B.J., Raghunth, G., Kuipers, B.: Parameter Learning from Stochastic Teachers and Stochastic Compulsive Liars. *IEEE Transaction Systems, Man, and Cybernetics* 36(4), 820–836 (2006)
3. Huang, D.H., Jiang, W.: A General CPL-AdS Methodology for Fixing Dynamic Parameters in Dual Environments. *IEEE Transaction Systems, Man, and Cybernetics* 42(5), 1489–1500 (2012)
4. Yazidi, A., Granmo, O.-C., Oommen, B.J., Goodwin, M.: A Hierarchical Learning Scheme for Solving the Stochastic Point Location Problem. In: Jiang, H., Ding, W., Ali, M., Wu, X. (eds.) *IEA/AIE 2012. LNCS (LNAI)*, vol. 7345, pp. 774–783. Springer, Heidelberg (2012)
5. Misra, S., Oommen, B.J.: GPSPA: A New Adaptive Algorithm for Maintaining Shortest Path Routing Trees in Stochastic Networks. *International Journal of Communication Systems* 17(10), 963–984 (2004)
6. Atlassis, A.F., Vasilakos, A.V.: The Use of Reinforcement Learning Algorithms in Traffic Control of High Speed Networks. In: *Advances in Computational Intelligence and Learning*, vol. 18, pp. 353–369 (2002)
7. Meybodi, M.R., Beigy, H.: New Learning Automata Based Algorithms for Adaptation of Backpropagation Algorithm Parameters. *International Journal of Neural Systems* 12(1), 45–67 (2002)

Application of GSO Algorithm to the Parameter Estimation of Option Pricing Model

Yuanbin Mo, Fuyong Liu, and Yanzhui Ma

Guangxi Key laboratory of hybrid computation and IC design analysis,
Nanning, 530006, China
moyuanbing@263.net

Abstract. The option pricing theory is an important part of modern finance, the Black-Scholes model is the cornerstone of modern option pricing theory. Therefore, an accurate estimate of the parameters of B-S model has important significance. Based on a new intelligent algorithm, artificial glowworm swarm optimization (GSO) algorithm, gives a generic method of B-S model parameter estimation. In this method, firstly, the parameters of B-S model is constructed artificial glowworm model, using the sum of square error between the predicted values and experimental values to design fitness function, and then through the fluorescein update and glowworms movement, to estimate the parameter of option pricing model. Based on many experiments, the results show that: the GSO for solving the above problem has good applicability and effectiveness, can obtain more accurate results, but the searching efficiency is gradually lowered if the ranges of parameter are widened.

Keywords: GSO, B-S model, parameter estimation, option pricing.

1 Introduction

Since the 1970s, the option pricing theory has been great developed so as to options trading has an important position in financial derivatives markets. In the process of the development of the option pricing theory, Black and Scholes [1] published a groundbreaking paper about option pricing in 1973, In this paper, they used the theory of stochastic differential equations and risk-free investment deduced the famous Black-Scholes model under some prerequisite conditions like that efficient markets and stock prices satisfy the geometric Brownian motion. This model is a milestone in the history of the development of option pricing, it laid a solid foundation for the pricing of option and even other for pricing rights and interests, make the option pricing that originally empty has the basis in theory.

The option price in Black-Scholes option pricing model depends on the following five parameters: Tagged asset market price S , Strike price E , maturity deadline $T-t$, Risk-free interest rate r , Tagged asset price volatility σ (the standard deviation of the return on assets). Among these parameters, the first three are predetermined value that easy access, but the risk-free interest rate and the tagged asset prices volatility are

required estimates by certain computational. Therefore, the inversion on the parameters of the option pricing model is carried out to determine r and σ .

Parameter inversion problem including the solution of the optimization models that option pricing parameters as variables and the solution of the option pricing simulation model. Binary tree method, Monte Carlo simulation, finite difference method [2] [9] and neural network method [3], etc, can be used to solve the option pricing simulation model. There are two broad categories optimization methods in general: the numerical method and intelligent method. The traditional numerical method has simplex method, the steepest descent method, conjugate gradient method, gauss-Newton method and variable metric method, etc. Option pricing problem is nonlinear, there may be multiple local optimal solutions, and the traditional numerical method largely depends on parameter initial value, and easy to fall into local optimum.

In recent years, various new bionic intelligent algorithms have been proposed with the development of computational intelligence technology, such as particle swarm optimization, simulated annealing algorithm, genetic algorithm, etc. However due to the different form and complexity of the various functions, therefore, each algorithm has its own advantages and disadvantages in the optimization problem. Glowworm Swarm Optimization (GSO) is a new type of swarm intelligent algorithm [4-5], in recent years; this algorithm aroused people's attention in the field of computational intelligence, and gradually become a hot area of research. With the deepening of the research, the algorithm has been successfully applied to the sensor noise test [6] and simulated robot [7], etc. The algorithm has the advantages of capture extreme value with high speed, capture with high efficiency and has strong universality, etc. This paper uses artificial firefly algorithm to establish the algorithm model of parameter estimation with the B-S option pricing model as the research object, the experimental result show that the effect of artificial firefly algorithm applied in B-S option pricing model parameters estimation problem is very good.

2 The Model of Black-Scholes

The basic assumption of the Black-Scholes model[8]: (1) allowing short selling securities in order to make full use of the proceeds;(2) there is no risk-free arbitrage opportunities;(3) there is no transaction costs and taxes, income securities are entirely separable;(4) there is no dividend or bonus in the validity period of the derivative securities;(5) during the validity period of derivative securities, the risk-free interest rate is constant; (6) the price of the underlying asset is a continuous variation, there is no jump or intermittent changes; (7) the evolution of stock price follows geometric Brownian motion, that is $dS = \mu Sdt + \sigma Sdw$, where, μ is expected return (constant), σ is the volatility (constant), $dw = \varepsilon\sqrt{dt}$ is a wiener process, ε obey the standard normal distribution.

Set S is stock price, r is the risk-free interest rate, E is the strike price of option, T is the maturity deadline of option, t is the current time, f is the option price and set up $f = f(S, t)$, deduce the Black - Scholes partial differential equations.

$$\frac{\partial f}{\partial t} + rS \frac{\partial f}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} = rf \tag{1}$$

Using the boundary conditions $f = \max\{S - E, 0\}$, get the price formula of the European call option

$$f = SN(d_1) - Ee^{-r(T-t)}N(d_2) \tag{2}$$

Where

$$d_1 = \frac{\ln(S/E) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \tag{3}$$

$$d_2 = \frac{\ln(S/E) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t} \tag{4}$$

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt \tag{5}$$

3 Artificial Glowworm Swarm Optimization Algorithm

Artificial glowworm swarm optimization algorithm is a new type of swarm intelligent algorithm that inspired by fireflies attract mating partners through flashing or the foraging behavior in nature. In the basic glowworm swarm optimization algorithm, first, initialize a group of fireflies in the solution space of the objective function randomly, every firefly with the same luciferin initial values, the fluorescence intensity of the firefly is proportional to luciferin value, and luciferin value is closely related to the fitness values of the landscape in the process of sports, the higher value of their corresponding luciferin, the stronger attraction to other fireflies within the scope of neighborhood. Neighborhood range called decision domain in GSO, its size is decided by its neighborhood radius r_d . In the movement of fireflies, every firefly goes forward to its neighbor fireflies within the neighborhood range in a certain probability. Fireflies j will become the neighbors of fireflies i , it must satisfy that fireflies j is in the neighborhood decision domain of fireflies i and the luciferin value of j is higher than i . Through the constant motion of the fireflies, the more fireflies finally gather around the fireflies that have the higher fitness value.

In GSO, each iteration consists of two stages: the first stage is luciferin update; the second stage is the movement of the firefly.

Luciferin update stage: at this stage every firefly updates luciferin according to (6)

$$l_i(t) = (1 - \rho)l_i(t-1) + \gamma J(x_i(t)) \tag{6}$$

Where, $l_i(t)$ is the luciferin value of the t -th generation of i -th firefly, $\rho \in (0,1)$ is the parameter of control luciferin value, γ is the parameter of evaluation function value, $J(x)$ is fitness evaluation function.

The movement stage of the firefly: at this stage firefly i choose its neighbor j within the neighborhood range and go forward to it in a certain probability, the formula of probability selection is as shown (7). The position of next time of firefly i is determined by (8), in the end of movement stage, every firefly updates decision domain according to (9).

The formula of path probability selection:

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)} \tag{7}$$

The formula of position update:

$$x_i(t+1) = x_i(t) + s * \left(\frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|} \right) \tag{8}$$

Where, $x_i(t) \in R^m$ is the position of firefly i in m dimension real space, $\|\bullet\|$ is the standard Euclidean distance metric operator, $s (> 0)$ is moving step length.

The formula of decision domain update:

$$r_d^i(t+1) = \min \left\{ r_s, \max \left\{ 0, r_d^i(t) + \beta \left(n_t - |N_i(t)| \right) \right\} \right\} \tag{9}$$

Where, β is a proportionality constant, n_t is the parameter that control the fireflies number within the neighborhood range, $|N_i(t)|$ is the number of fireflies neighbor within the neighborhood range.

4 Build the Fitness Evaluation Function

Using the finite difference method [9] to get a set of simulation experimental value through preset a set of parameters $S, E, T-t, r, \sigma$, get the forecasting value by intelligent algorithm, taking the sum of error square of experimental and forecasting value as the objective optimization function, specific mathematical form such as (10)

$$J = \sum_{i=1}^N [y_{pred}(i) - y_{exp}(i)]^2 \tag{10}$$

Where, N is the number of data we get from each experiment, y_{exp} is the experimental value of artificial simulation, and y_{pred} is the forecasting value of algorithm.

5 Example Analysis

5.1 Sources of Date

This paper quotes the data of literature [9], GSO is applied to solve the parameters of Black-Scholes option pricing model. The values of five factors affecting option prices are given in this literature(the stock price S , strike price E , risk-free interest rate r , the time of maturity date T ,volatility σ), and give out the option price of market observations and the option prices that use the finite difference method to solve, a total of 30 groups. Specific see [9]

5.2 Experimental Conditions

The numerical experiment uses Matlab language. In the experiment, the maximum number of iterations is 500, the numbers of particles respectively take 10, 20, 50, 100, 150 and 200, and the initial range of option parameters that will be estimated is: the minimum values are all zero, the maximum values are respectively 0.2, 0.4, 0.6, 0.8 and 1.0 of the true values of parameters.

Table 1. The relation between algorithm iterations and the number of fireflies and the initial range of parameter

Number of fireflies	Initial range of parameter	Iterations	Number of fireflies	Initial range of parameter	Iterations
10	[0 0.2]	130	100	[0 0.2]	92
10	[0 0.4]	155	100	[0 0.4]	103
10	[0 0.6]	198 **	100	[0 0.6]	110
10	[0 0.8]	254 **	100	[0 0.8]	114
10	[0 1.0]	269 **	100	[0 1.0]	120
20	[0 0.2]	118	150	[0 0.2]	87
20	[0 0.4]	129	150	[0 0.4]	95
20	[0 0.6]	150	150	[0 0.6]	104
20	[0 0.8]	195	150	[0 0.8]	109
20	[0 1.0]	214	150	[0 1.0]	111
50	[0 0.2]	101	200	[0 0.2]	74
50	[0 0.4]	114	200	[0 0.4]	91
50	[0 0.6]	119	200	[0 0.6]	97
50	[0 0.8]	125	200	[0 0.8]	103
50	[0 1.0]	133	200	[0 1.0]	103

5.3 Experimental Results

The iteration times numerical experiment results under the condition of different number of fireflies and different initial range of option parameters that will be estimated are shown in table 1. It is necessary to explain that: due to the initial

position of fireflies are given randomly, there may appear different results in the repeat operation, but the results can reflect a certain operation time consuming trend. Table 1 selected the calculation case of parameter r , the number of iterations is take the average number of 20 times calculate iterations, **mean the condition that it still does not meet the accuracy requirements after iterating 500 times in 20 times computing.

Table 2 respectively make a list of the experimental values and the fitting values of algorithm that are given by r and σ , the population size is 100, the algorithm iterations is 150, the initial range of parameter is [0 1.0], the fitting value of algorithm is the optimal fitting values after the program runs 20 times.

Table 2. The experimental values and the fitting values of algorithm of r and σ

Sequence number	r fitting values of algorithm	r fitting values of algorithm	r fitting values of algorithm	r fitting values of algorithm
1	0.0771	0.077101	0.161633	0.1616329
2	0.0771	0.077099	0.161633	0.1616322
3	0.0771	0.077101	0.161633	0.1616334
4	0.0771	0.077101	0.161633	0.1616328
5	0.0771	0.077101	0.161633	0.1616324
6	0.0771	0.077099	0.161633	0.1616336
7	0.0771	0.077102	0.161633	0.1616333
8	0.0777	0.077707	0.166284	0.1662849
9	0.0777	0.077698	0.166284	0.1662848
10	0.0777	0.077703	0.166284	0.1662843
11	0.0777	0.077699	0.166284	0.1662841
12	0.0777	0.077701	0.166284	0.1662844
13	0.0777	0.077701	0.166284	0.1662845
14	0.0777	0.077699	0.159941	0.1599411
15	0.0777	0.077702	0.159941	0.1599406
16	0.0777	0.077702	0.159941	0.1599400
17	0.0777	0.077698	0.159941	0.1599402
18	0.0777	0.077698	0.159941	0.1599415
19	0.0777	0.077699	0.158642	0.1586420
20	0.0777	0.077701	0.158642	0.1586417
21	0.0777	0.077701	0.158642	0.1586415
22	0.0777	0.077700	0.158642	0.1586422
23	0.0777	0.077703	0.158642	0.1586424
24	0.0777	0.077699	0.158642	0.1586411
25	0.0777	0.077701	0.158642	0.1586428
26	0.0777	0.077702	0.136054	0.1360475
27	0.0777	0.077698	0.136054	0.1360544
28	0.0777	0.077699	0.136054	0.1360547
29	0.0777	0.077699	0.136054	0.1360533
30	0.0777	0.077699	0.136054	0.1360543

5.4 Results Analysis

It is clearly observed that the number of fireflies and the initial range of option parameters that will be estimated have obvious influence on the process of calculation from table 1. In terms of overall trend, the greater fireflies number, the less number of iterations is needed, and the greater initial range of parameters, the more number of iterations are needed. Under the experimental conditions, when the number of particles is too small and the initial range of option parameters that will be estimated is larger, there will be an operational failure or misconvergence. The above phenomenon indicates that, when we use artificial glowworm swarm optimization algorithm, we had better employ more number of fireflies to operation, and analysts should according to their own professional knowledge, preliminary estimate a relatively appropriate possible values range of option parameters, to ensure the convergence of the operation process and can get the result quickly.

It is observed that the error of the fitting values of algorithm and parameter experimental values is very small from table 2. It shows that artificial glowworm swarm optimization algorithm on solving B-S model parameters problem has good applicability and effectiveness, can obtain more accurate results.

6 Conclusion

In this paper, a new type of swarm intelligent algorithm--artificial glowworm swarm optimization algorithm is applied to solve the problem of B-S option pricing model parameters estimation. Research results show that: artificial glowworm swarm optimization algorithm can obtain more satisfactory results; the number of fireflies and the variation range of option parameters have obvious influence on the convergence of the algorithm and the number of iterations. Therefore, the convergence theory of the algorithm and how to reduce or get rid of the influence of the variable range are the problems that we need to do further research.

Acknowledgments. The authors are very grateful to the referees for their valuable comments and suggestions. These works are supported by Guangxi key laboratory of hybrid computation and IC design analysis open fund (2012HC08), and China Postdoctoral Foundation (2012M511711), and Guangxi Department of Education Item (201204LX082) and item of Guangxi University for Nationalities (2011MDYB030).

References

1. Black, F., Scholes, M.: The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3), 637–654 (1973)
2. Jiang, L.S.: *The mathematical model and methods of option pricing*. Higher Education Press, Beijing (2003)

3. Wu, L.Y., Ma, W.W.: Real Options Pricing Based on Artificial Neural Network. *Journal of Wuhan University of Science and Technology* 28(1), 80–83 (2004)
4. Krishnanand, K.N., Ghose, D.: Glowworm Swarm Optimization: A New method for Optimizing Multi-modal functions. *International Journal of Computational Intelligence Studies* 1(1), 93–119 (2009)
5. Krishnanand, K.N.: Glowworm Swarm Optimization: A Multimodal Function Optimization Paradigm with Applications to Multiple Signal Source Localization Tasks. Indian Institute of Science, [S. l.] (2007)
6. Krishnanand, K.N., Ghose, D.: A Glowworm Swarm Optimization Based Multi-robot System for Signal Source Localization. [s. n.], Berlin (2009)
7. Krishnanand, K.N., Ghose, D.: Chasing Multiple Mobile Signal Sources: A Glowworm Swarm Optimization Approach. In: *Proc. of the 3rd Indian International Conference on Artificial Intelligence*. IEEE Press, [S. l.] (2007)
8. Chen, X.H.: *Micro finance*. Shanghai University of Finance and Economics Press, Shanghai (2006)
9. Zhao, X.: Parameter estimation of option pricing model based on intelligent optimization algorithms. Master's degree paper. Jiangnan University, Wuxi (2011)

Heuristics for Hub Location Problems with Alternative Capacity Levels and Allocation Constraints

Jeng-Fung Chen

Department of Industrial Engineering & Systems Management, Feng Chia University
j fchen@fcu.edu.tw

Abstract. We consider the hub location problem in which the capacity of a hub restricts the amount of traffic transiting through the hub, the capacity of a hub is a decision variable, and each non-hub can only be allocated to a certain number of hubs. The objective is to determine the number of hubs, to locate the hubs, to determine the capacity level of each hub, to allocate the non-hubs to the hubs, and to determine the path for each origin-destination pair such that the total cost is minimized. Two heuristics based on the threshold accepting method are proposed to resolve this type of hub location problems. Computational characteristics of the proposed heuristics are evaluated through computational experiments using the modified CAB and AP data sets.

Keywords: allocation constraint, alternative capacity, capacitated hub location problem, hub-and-spoke network.

1 Introduction

Hub location problems are decision problems regarding locating hub facilities, allocating the non-hubs (i.e., spokes) to the hubs, and determining the path for each origin-destination pair. In many telecommunication and transportation networks, connecting all pairs of nodes by direct links is costly. Hubs therefore are established to serve as transshipment points. The traffic from the same origin with different destinations is collected on the allocated hub(s), where it is regrouped with other traffic from different origins and leaves the hub(s) either to its ultimate destinations or to other hubs. Hubs are fully interconnected and the unit transportation costs on all inter-hub links are discounted by a factor to reflect economies of scale on the inter-hub links. Direct traffic between non-hubs is costly and is usually not allowed. Since much research has indicated that implementing hub-and-spoke networks can reduce the overall cost, there has been a rapid growth in the application of hub-and-spoke networks. Typical applications can be found in [1-2].

There are several versions of hub location problems [3]. In the p -hub median problem, the number of hubs p is predetermined. The objective is to locate the hubs, to allocate the non-hubs to the hubs, and to determine the path for each origin-destination pair such that the total transportation cost is minimized, which includes (1) the collection cost incurred during the transportation from the non-hubs to their allocated hub(s); (2) the transfer cost incurred during the transportation between hubs;

and (3) the distribution cost incurred during the transportation from the allocated hub(s) to the non-hubs. The uncapacitated hub location problem is a different type of hub location problem, in which the number of hubs is not given *a priori* and a cost for installing a hub is included in the objective function. The capacitated hub location problem is another version of hub location problem, in which capacity constraints are enforced on the hubs and/or the arcs. Two types of allocation are commonly found in the literature. The multiple allocations allow each non-hub to be routed to all the hubs. In the single allocation, each non-hub is allocated to only a single hub.

Since the 1990s, capacitated hub location problems have attracted a lot of research attention. For example, Campbell [4] presented a mixed-integer linear program for the capacitated multiple allocation hub location problem. Carello *et al.* [5] proposed different formulations and a local search approach. Rodriguez-Martin and Salazar-Gonzalez [6] proposed two branch-and-bound methods and a heuristic. For the capacitated single allocation hub location problem, Ernst and Krishnamoorthy [7] developed two heuristics and a branch-and-bound solution method to solve the problem. Randall [8] developed ant colony-based heuristics to solve the problem.

Although hub location problems have been well studied, very little research has considered that each non-hub can only be connected to a few hubs, perhaps 2 or 3, as commonly found in many practical applications. In this research, we deal with this type of hub location problems. Furthermore, in order to reflect practical applications of planning hub-and-spoke networks, the costs for installing hubs are regarded as discrete functions of their capacity levels. Yaman [9] studied the allocation strategies. However, he mainly focused on the *p*-hub median problem. Correia *et al.* [10] considered that the capacity of the hubs is part of the decision making process. However, they only considered single allocation case. In this paper, two heuristics based on the threshold accepting method [11] and improvement procedures are proposed to resolve this problem. Computational characteristics of the proposed heuristics are evaluated through computational experiments using the modified CAB and AP data sets.

2 Mathematic Model

Let W_{ij} be the traffic from origin i to destination j ; χ be the unit collection cost; α be the unit transportation cost on the inter-hub links; δ be the unit distribution cost; c_{ik} be the distance between nodes i and k ; $X_{ijkm} = 1$ if the traffic from origin i to destination j routed via hubs k and m in that order, and 0 otherwise; $F_{k\ell}$ be the fixed cost of establishing hub k at level ℓ ; $\Gamma_{k\ell}$ be the capacity of hub k at level ℓ ; $Z_{kk\ell} = 1$ if node k is a hub with capacity at level ℓ , and 0 otherwise; $Z_{ik\ell} = 1$ if node i can be allocated to hub k , and 0 otherwise; h be the number of hubs that each non-hub can be allocated to. This hub location problem can be formulated as

$$\text{Min} \quad \sum_i \sum_j \sum_k \sum_m (\chi c_{ik} + \alpha c_{km} + \delta c_{mj}) W_{ij} X_{ijkm} + \sum_k \sum_\ell F_{k\ell} Z_{kk\ell}$$

$$\text{S.T.} \quad \sum_k \sum_m X_{ijkm} = 1 \qquad \forall i, j \qquad (1)$$

$$\sum_\ell Z_{kk\ell} \leq 1 \qquad \forall k \qquad (2)$$

$$\sum_k \sum_\ell Z_{ik\ell} \leq h \qquad \forall i \qquad (3)$$

$$Z_{ikt} - Z_{kk\ell} \leq 0 \quad \forall i, k, \ell \quad (4)$$

$$\sum_m X_{ijkm} - \sum_\ell Z_{ik\ell} \leq 0 \quad \forall i, j, k \quad (5)$$

$$\sum_k X_{ijkm} - \sum_m Z_{jm\ell} \leq 0 \quad \forall i, j, m \quad (6)$$

$$\sum_i \sum_j \sum_m W_{ij} X_{ijkm} + \sum_i \sum_j \sum_{m \neq k} W_{ij} X_{ijkm} \leq \sum_\ell \Gamma_{k\ell} Z_{kk\ell} \quad \forall k \quad (7)$$

$$Z_{ikt} \in \{0, 1\} \quad \forall i, j, \ell$$

$$X_{ijkm} \in \{0, 1\} \quad \forall i, j, k, m$$

The objective is to minimize the total costs of collection, transfer, distribution, and installing hubs. Constraint (1) requires the traffic between every origin-destination pair must be routed via some hub pair. Constraint (2) assures each hub can only be installed at one capacity level. Constraint (3) ensures that each non-hub can only be allocated to h hubs. Constraint (4) implies that a node should be established as a hub before any non-hub is allocated to it. Constraints (5) and (6) enforce $\sum_\ell Z_{ik\ell}$ ($\sum_\ell Z_{jm\ell}$) to be one if there is any traffic between node i (j) and hub k (m). The capacity constraint of each hub is ensured by Constraint (7), considering both collection and transfer.

3 Heuristics

The proposed heuristics, based on threshold accepting method, are detailed in this section. First, the threshold accepting method is briefly described.

3.1 Threshold Accepting Method

The threshold accepting method (TA) was introduced by [11]. TA accepts a neighborhood solution (Y) if its solution value ($V(Y)$) is not worse than the best-known solution value plus a gradual lowered *THRESHOLD*.

Threshold accepting method for minimization

generate an initial solution and choose an initial THRESHOLD (T) > 0
Repeat: generate a neighborhood that is a perturbation of the current solution
IF $V(Y) < V(X) + T$
 THEN accept the neighborhood solution (i.e., set $X = Y$)
 IF no improvement over the solution for a number of iterations
 THEN lower T
IF the stop criterion is met
 THEN stop
GOTO Repeat

3.2 Heuristic TA_A

Heuristic TA_A consists of four levels: the first level is to determine the number of hubs; the second level is to locate the hub facilities; the third level is to allocate the non-hubs to the hubs; and the fourth level is to determine the path for every origin-destination pair and the capacity level of each hub.

3.2.1 Determining the Number of Hubs

To obtain the optimal/near optimal solution efficiently, we can employ the lower bound for the number of hubs and the condition for the optimal number of hubs. A pseudo-lower bound (p'') can be set to be $h+2$. Furthermore, based on Chen [12], the optimal number of hubs (p) is equal to $p'-1$ if the reduction of the transportation cost is greater than the increase in establishing cost when $p=p'-1$ and no greater than the increase in establishing cost when $p=p'$.

3.2.2 Locating Hub Facilities

To choose good hub locations, the components in the objective function need to be considered. Therefore, the index $I_i = (W_i / \sum_j W_j) + (1 - D_i / \sum_j D_j) + (\Omega_i / \sum_j \Omega_j)$ of each node is taken into account when locating hubs, where W_i is the total flow to and from node i , D_i is the total distance between node i and the other nodes, and Ω_i is the average capacity of node i divided by the average cost for installing hub i .

Generating initial hub locations

Let o with coordinate $((\sum_i x_i m_i) / \sum_i m_i, (\sum_i y_i m_i) / \sum_i m_i)$ be the center of mass of the system, in which m_i is the mass of node i and is set to be $\sum_j (W_{ji} + W_{ij}) / 2$. When generating initial hub locations, regard o as the origin and all the nodes are divided into p groups. For instance, if there are two hubs, group 1 comprises all the nodes within the region formed by the 0° and 180° rays and group 2 comprises all the nodes within the region formed by the 180° and 360° rays. Then, a node in group n to be selected as a hub is based on probability $p_i = I_i / \sum_{j \in N_n} I_j$, where N_n contains all the nodes in group n . Then, the capacity of each selected hub is set at the second level.

Generating neighborhood solutions

We used exchange procedure to generate neighborhood solutions. In this procedure, each of the hubs is alternatively exchanged with one non-hub (i), each randomly selected based on probability $p_i = I_i / \sum_{j \in NH} I_j$ with the restriction that the distance between them (the one to replace and the one to be replaced) cannot exceed $D/2$, where N is the set containing all the nodes, H is the current hub set, and D is the maximum distance between any two nodes in N . The procedures in Sections 3.2.3 and 3.2.4 are applied to determine the path of every origin-destination pair. The exchange obtaining the best result is chosen. A tabu list is employed to record the hub sets searched. If any exchange may result in a hub set searched, that exchange is prohibited.

3.2.3 Allocating the Non-hubs to the Hubs

We first relax this problem to the multiple allocation case and apply the procedures described in Section 3.2.4 to determine the path for every origin-destination pair and the capacity level of each hub. If there are no non-hubs violating the allocation constraints, the obtained paths are optimal for the given hub set. Otherwise, record the non-hubs violating the allocation constraints and each of their connected hubs. Let M_{i1} be the hub such that the maximum flow to and from non-hub i occurs on the link between non-hub i and hub M_{i1} , M_{i2} be the hub such that the second-maximum flow to and from non-hub i occurs on the link between non-hub i and hub M_{i2} , and so forth.

Then, for every non-hub only the links with hubs $M_{i1}, M_{i2}, \dots, M_{ih}$ remain; the other links are deleted and the paths for the associated origin-destination pairs are re-determined. The obtained solution is then improved by the following procedure.

Improvement procedure

Step1: For the non-hubs violating the allocation constraints, starting from the non-hub with the least total flows, the links to the recorded alternative hubs are tried (with M_{i1} or M_{i2} remains connected) and the paths for the violated origin-destination pairs are re-determined. The alternative links and capacity levels that lead to the best result are selected. If all the non-hubs are tried, proceed to step 2.

Step2: If the solution is improved, return to step 1. Otherwise, stop.

3.2.4 Determining the Path for Each Origin-Destination Pair

We first used the shortest path algorithm under capacity constraints to determine the path for every origin-destination pair. The obtained path for every origin-destination pair then is refined by the following improvement procedure.

Modified shortest-path procedure.

Step 0: Let C^1_{ij} be the unit cost of pair ij if it is routed via its shortest path, C^2_{ij} be the unit cost of pair ij if it is routed via its second-shortest path, and $O^2_{ij} = C^2_{ij} - C^1_{ij}$. Set $Count=0$ and $Index(k)=0$ for every hub k .

Step 1: The path for the pair with the largest O^2_{ij} is determined first. If there is any pair not routed via its shortest path due to capacity constraints, record its current unit cost, C^c_{ij} , and the associated hub(s) that results in this situation.

Step 2: Let $Q(k)$ be the set of pairs not routed hub k but their shortest paths must be routed via hub k and $R(k)$ be the set of pairs currently routed hub k . Compute $O^c_{ij} = C^c_{ij} - C^1_{ij}$ for every ij in set $Q(k)$ and $T(k) = \sum_{ij \in Q(k)} W_{ij} O^c_{ij}$ for every hub k .

Step 3: If $Index(k)=1$ or $T(k)=0$ for every k , stop. Otherwise, select hub $k^* = \text{argmax}_{Index(k)=0} \{T(k)\}$.

Step 3-1: Sort pairs in set $Q(k^*)$ in decreasing order of O^c_{ij} .

Step 3-2: Solve the knapsack problem to minimize the cost satisfying that the randomly selected pairs with larger O^c_{ij} in set $Q(k^*)$ can be routed via hub k^* if the selected pairs in set $R(k)$ are not routed via hub k^* . If the solution is improved, set $Index(k^*)=0$ for every k and return to step 2.

Step 4: Remove pair i^*j^* from set $Q(k^*)$. If $Count < 5$, set $Count=Count+1$ and return to step 3-2. Otherwise, set $Index(k^*)=1$ and return to step 3.

After determining the path for every origin-destination pair, a *post-procedure* is applied to improve the capacity level decision of each hub.

Post-procedure

Step 1: If there is any hub with too much capacity left (with too little capacity left) and its capacity level can be decreased (increased), decrease (increase) its capacity level and proceed to step 2. Otherwise, stop.

Step 2: Perform the *modified shortest-path procedure* and return to step 1.

3.2.5 Heuristic TA_A

Figure 1 shows the structure of Heuristic TA_A. TA_A is detailed as follows.

Step 0: Initialize *Counter1*, *Counter2*, and *THRESHOLD (T)* and set *Counter3*.

Step 1: Set $p=p''$ (to set the initial number of hubs to be the pseudo-lower bound) and the total cost to be infinite for $p=p''-1$.

Step 2: Generate the initial hub locations, their capacity levels, and apply the procedures for allocating the non-hubs to the hubs and determining the path for every origin-destination pair.

Step 3: Apply the exchange procedure and the procedures for allocating the non-hubs to the hubs, determining the path for every pair, and adjusting the capacity levels for the hubs. Apply the TA module to determine whether this solution is accepted. Set $Counter2=Counter2+1$ and update *T*. If the best-known solution (for the given number of hubs) is improved, set $Counter1=0$. Otherwise, set $Counter1=Counter1+1$. If a *Counter3* number of neighborhood searches have been applied without improvement over the best-known solution (for the given number of hubs), proceed to step 4.

Step 4: If $p < p''$, proceed to step 4-2.

Step 4-1: If the resulted reduction of the transportation cost is greater than the increase in installing cost, set $p=p+1$ (increase the number of hubs), reinitialize *Counter1*, *Counter2*, and *T*, and return to step 2. If $p=p''+1$, set $p=p''-1$ (to decrease the number of hubs), reinitialize *Counter1*, *Counter2*, and *T*, and return to step 2. Otherwise, stop.

Step 4-2: If the resulted increase in the transportation cost is greater than the decrease in installing cost, stop. Otherwise, set $p=p-1$, reinitialize *Counter1*, *Counter2*, and *T*, and return to step 2.

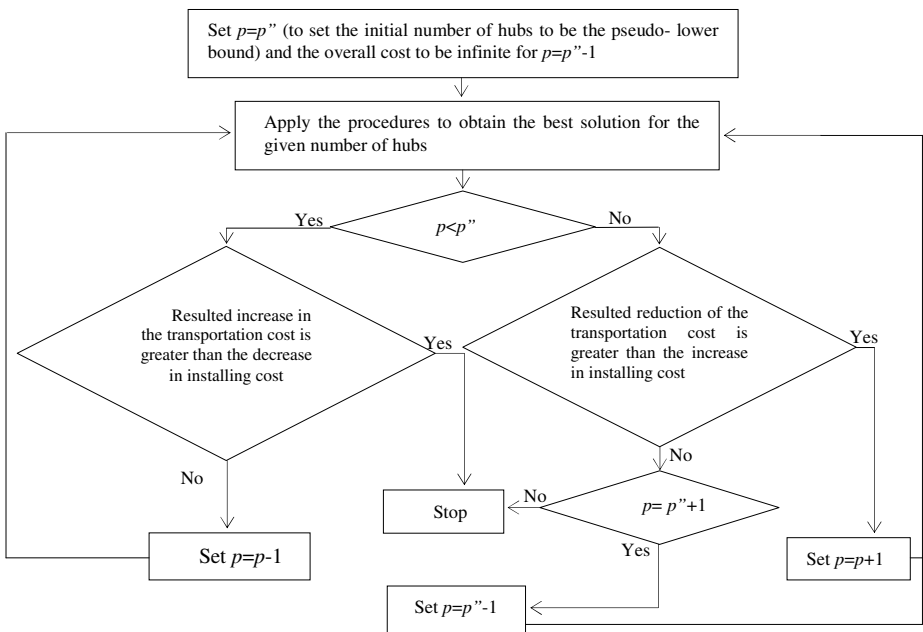


Fig. 1. Structure of Heuristic TA_A

3.3 Heuristic TA_B

Heuristic TA_B is very similar to TA_A. The main difference is in the procedure to allocate the non-hubs to the hubs. In TA_B, each non-hub is first allocated to its h nearest hubs. The procedures described in Section 3.2.4 are applied to determine the path of every origin-destination pair and the capacity level of each hub. Then, the obtained solution is improved by the following procedure.

Improvement procedure

Step1: Starting from the non-hub with the least total flows, the allocations to the alternative hubs are tried and the path for every origin-destination pair is re-determined. The alternative links and capacity levels that lead to the best solution are selected. If all the non-hubs have been tried, proceed to step 2.

Step 2: If the solution is improved, return to step 1. Otherwise, stop.

4 Computational Experiments

The CAB and AP data sets were modified to evaluate the performance of TA_A and TA_B. Since there are no related methods applicable in the literature, the solutions of TA_A and TA_B were compared with the optimal solutions obtained by Cplex. Both heuristics were coded in C++ and the experiments were performed on the same PC.

4.1 Data Sets

The CAB data set was introduced by [13]. For the modified CAB data set, the unit collection cost (γ) and the unit distribution cost (δ) were both set to be 1, while the unit cost on the inter-hub links (α) was set to be 0.5 or 0.6. It was assumed there are three capacity levels for every node. For the third level, the capacity of node i (F_i) was set to be $3(n/p+3d_iO_i/(5d_mO_m))O_i$, where n was the problem size, d_i was the distance from node i to the center of mass of the system, O_i is the total collection at node i , $d_m = \max\{d_i\}$, $O_m = \max\{O_i\}$, p was set at 9 or 10. For c type problems, the cost for installing hub i was set to be $F_i^c = f_0(1-3(F_i+O_i)/(5(\Gamma_m+O_m)))$, where $\Gamma_m = \max\{\Gamma_i\}$, $f_0 = (\sum_i \sum_j (\gamma C_{ie} + \delta C_{ej}) W_{ij} - \sum_i \sum_j \alpha C_{ej} W_{ij})/p$, e was the node nearest to the center of mass, and C_{ie} was the distance between nodes i and e ; for d type problems, the cost for installing hub i was set to be $F_i^d = f_0(1-3d_i/(5d_m))$. For the second level, the capacity and installing cost of each node were set to be 2/3 and 4/5 of the third level; for the first level, the capacity and installing cost of each node were set to be 1/3 and 3/5 of the third level. The number of hubs that each non-hub can be allocated to was set to be 3.

The AP data set was introduced by [14]. For each problem size, two types of installing costs were considered: tight (T) and loose (L). Two types of capacities were also considered: tight and loose. In order to use this data set to test this type of hub location problems, the unit collection cost, unit transportation cost on the inter-hub links, and the unit distribution cost were set to be 5.5, 3.0, and 4.5, respectively. For the third level, the capacity and installing cost of each node were set to be 200% and 0.75x55% of the original; for the second level, the capacity and installing cost of each node were set to be 170% and 0.7x55% of the original; for the first level, the capacity

and installing cost of each node were set to be 140% and 0.65x55% of the original. The number of hubs that each non-hub can be allocated to was set to be 2 or 3.

4.2 Computational Results

The reported solutions were based on running both heuristics ten times each with a different random seed, given that *Counter1* was set at $\lceil \sqrt{n} \rceil$.

Table 1. Computational results of the modified CAB data set

<i>n-t-p-α</i>	Optimal cost by Cplex	TA_A				TA_B			
		Min gap	Max gap	Avg gap	CPU time	Min gap	Max gap	Avg gap	CPU Time
10-c9-0.5	647.94	0.00%	0.00%	0.00%	0.02	0.00%	0.00%	0.00%	0.05
10-c10-0.5	631.18	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.05
10-d9-0.5	630.50	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.06
10-d10-0.5	609.78	0.00%	0.00%	0.00%	0.04	0.00%	0.00%	0.00%	0.07
10-c9-0.6	672.43	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.06
10-c10-0.6	658.77	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.08
10-d9-0.6	655.57	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.09
10-d10-0.6	637.65	0.00%	0.00%	0.00%	0.04	0.00%	0.00%	0.00%	0.08
15-c9-0.5	969.74	0.00%	0.00%	0.00%	0.13	0.00%	0.00%	0.00%	0.43
15-c10-0.5	977.73	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.45
15-d9-0.5	980.65	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.46
15-d10-0.5	983.00	0.00%	0.00%	0.00%	0.18	0.00%	0.00%	0.00%	0.47
15-c9-0.6	1003.80	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.49
15-c10-0.6	1010.78	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.47
15-d9-0.6	1013.66	0.00%	0.00%	0.00%	0.19	0.00%	0.00%	0.00%	0.46
15-d10-0.6	1015.28	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.53
20-c9-0.5	905.77	0.00%	0.00%	0.00%	0.70	0.00%	0.00%	0.00%	1.60
20-c10-0.5	882.14	0.00%	0.00%	0.00%	0.71	0.00%	0.00%	0.00%	1.62
20-d9-0.5	920.05	0.00%	0.00%	0.00%	0.72	0.00%	0.00%	0.00%	1.62
20-d10-0.5	894.73	0.00%	0.00%	0.00%	0.71	0.00%	0.00%	0.00%	1.63
20-c9-0.6	942.96	0.00%	0.00%	0.00%	0.78	0.00%	0.00%	0.00%	1.62
20-c10-0.6	921.68	0.00%	0.00%	0.00%	0.73	0.00%	0.00%	0.00%	1.64
20-d9-0.6	955.84	0.00%	0.00%	0.00%	0.72	0.00%	0.00%	0.00%	1.74
20-d10-0.6	933.07	0.00%	0.00%	0.00%	0.79	0.00%	0.00%	0.00%	1.75
25-c9-0.5	1046.16	0.00%	0.00%	0.00%	1.37	0.00%	0.00%	0.00%	3.14
25-c10-0.5	1038.78	0.00%	0.00%	0.00%	1.46	0.00%	0.00%	0.00%	3.14
25-d9-0.5	1052.26	0.00%	0.00%	0.00%	1.36	0.00%	0.03%	0.01%	3.15
25-d10-0.5	1031.70	0.00%	0.02%	0.01%	1.38	0.00%	0.04%	0.01%	3.17
25-c9-0.6	1078.80	0.00%	0.00%	0.00%	1.48	0.00%	0.00%	0.00%	3.24
25-c10-0.6	1073.62	0.00%	0.00%	0.00%	1.37	0.00%	0.00%	0.00%	3.15
25-d9-0.6	1089.87	0.00%	0.00%	0.00%	1.70	0.00%	0.03%	0.01%	3.30
25-d10-0.6	1068.02	0.00%	0.02%	0.01%	1.77	0.00%	0.05%	0.02%	3.29

4.2.1 Results of the Modified CAB Data Set

The computational results of the modified CAB problems are given in Table 1. According to the computational results, TA_A was able to consistently obtain the optimal solution in almost every run and TA_B was also capable of obtaining good

results. As for the runtime consumed, the required runtimes ranged from 0.02 to 3.30 CPU seconds. The maximum gap was within 0.05%. This indicates that both heuristics were able to solve the modified CAB problems efficiently and effectively.

Table 2. Computational results of the modified AP data set

h	n-t	Optimal cost by Cplex [*]	TA_A				TA_B			
			Min Gap	Max gap	Avg Gap	CPU time	Min gap	Max gap	Avg gap	CPU time
2	10LL	260161.58	0.00%	0.00%	0.00%	0.03	0.00%	0.00%	0.00%	0.04
	10LT	300157.18	0.00%	0.00%	0.00%	0.04	0.00%	0.00%	0.00%	0.05
	10TL	303599.62	0.00%	0.00%	0.00%	0.04	0.00%	0.00%	0.00%	0.05
	10TT	323358.78	0.00%	0.00%	0.00%	0.05	0.00%	0.00%	0.00%	0.06
	20LL	333720.89	0.00%	0.00%	0.00%	0.16	0.00%	0.00%	0.00%	0.21
	20LT	337101.61	0.00%	0.00%	0.00%	0.25	0.00%	0.00%	0.00%	0.33
	20TL	390042.21	0.00%	0.00%	0.00%	0.26	0.00%	0.00%	0.00%	0.34
	20TT	394444.43	0.00%	0.00%	0.00%	0.22	0.00%	0.00%	0.00%	0.28
	25LL	339699.26	0.00%	0.00%	0.00%	1.63	0.00%	0.00%	0.00%	2.03
	25LT	354068.76	0.00%	0.00%	0.00%	1.53	0.00%	0.00%	0.00%	1.90
	25TL	414166.43	0.00%	0.00%	0.00%	1.83	0.00%	0.00%	0.00%	2.30
	25TT	424072.29	0.00%	0.00%	0.00%	1.72	0.00%	0.02%	0.01%	2.15
	40LL	363798.49	0.00%	0.00%	0.00%	9.40	0.00%	0.01%	0.01%	11.75
	40LT	371729.96	0.00%	0.02%	0.01%	9.35	0.00%	0.03%	0.00%	11.69
	40TL	424443.76	0.00%	0.00%	0.00%	9.46	0.00%	0.02%	0.01%	11.82
	40TT	456211.10	0.00%	0.02%	0.01%	9.62	0.00%	0.04%	0.02%	12.03
3	10LL	260161.58	0.00%	0.00%	0.00%	0.02	0.00%	0.00	0.00%	0.05
	10LT	287688.65	0.00%	0.00%	0.00%	0.02	0.00%	0.00	0.00%	0.05
	10TL	303576.23	0.00%	0.00%	0.00%	0.03	0.00%	0.00	0.00%	0.06
	10TT	318258.25	0.00%	0.00%	0.00%	0.04	0.00%	0.00	0.00%	0.06
	20LL	331691.09	0.00%	0.00%	0.00%	0.23	0.00%	0.00	0.00%	0.41
	20LT	334708.88	0.00%	0.00%	0.00%	0.24	0.00%	0.00	0.00%	0.42
	20TL	388286.71	0.00%	0.00%	0.00%	0.25	0.00%	0.00	0.00%	0.43
	20TT	388864.55	0.00%	0.00%	0.00%	0.26	0.00%	0.00	0.00%	0.45
	25LL	337195.51	0.00%	0.00%	0.00%	1.36	0.00%	0.00	0.00%	2.38
	25LT	350226.48	0.00%	0.00%	0.00%	1.32	0.00%	0.00	0.00%	2.31
	25TL	410597.35	0.00%	0.00%	0.00%	1.51	0.00%	0.00	0.00%	2.66
	25TT	419314.09	0.00%	0.02%	0.01%	1.60	0.00%	0.03	0.01%	2.80
	40LL	360831.23	0.00%	0.02%	0.01%	8.53	0.00%	0.03	0.01%	14.93
	40LT	369568.12	0.00%	0.00%	0.00%	8.39	0.00%	0.04	0.01%	14.70
	40TL	422944.78	0.00%	0.02%	0.01%	8.82	0.00%	0.05	0.02%	15.43
	40TT	451145.08	0.00%	0.03%	0.01%	8.89	0.00%	0.05	0.02%	15.55

^{*}The indicated solutions for problems with 40 nodes were the best solution obtained by TA_A or TA_B.

4.2.2 Results of the Modified AP Data Set

Table 2 shows the computational results of the modified AP problems. Cplex failed to obtain the optimal solutions for the tested problems with 40 nodes, due to running out of memory. According to the computational results, TA_A and TA_B were capable of obtaining the optimal solutions for the small-scale modified AP problems (with 10-25 nodes) almost in every run. Although the solutions produced by TA_A and TA_B became more inconsistent when the number of nodes increased, the maximum gap was within 0.05%. This indicates that both heuristics were able to produce good solutions for the modified AP problems. The required runtimes ranged from 0.02 to 15.55

CPU seconds. TA_A performed better than TA_B not only on the solution quality but also on the runtime. Based on these computational results, the allocation procedure of TA_A can produce better results for this type of hub location problems.

5 Conclusions and Suggestions for Future Research

In this research, we have dealt with hub location problems with alternative capacity levels and allocation constraints. Two heuristics based on the threshold accepting method and improvement procedures have been proposed to resolve this type of hub location problems. Computational characteristics of the proposed heuristics have been evaluated through computational experiments using the modified CAB and AP data sets. Computational results have demonstrated that the proposed heuristics are capable of obtaining optimal solutions for small-scale problems. As for future research, it may be desirable to develop exact algorithms to obtain the optimal solutions.

References

1. Lee, Y., Lim, B., Park, J.: A hub location problem in designing digital data service networks: Lagrangian relaxation approach. *Location Science* 4, 185–194 (1996)
2. Kubly, M., Gray, R.: Hub network design problem with stopovers and feeders: case of Federal Express. *Transportation Research* 27, 1–12 (1993)
3. Campbell, J., O’Kelly, M.E.: Twenty-five years of hub location research. *Transportation Science* 46, 153–169 (2012)
4. Campbell, J.: Integer programming formulations of discrete hub location problems. *European Journal of Operational Research* 72, 387–405 (1994)
5. Carello, G., Della Croce, F., Ghirardi, M., Tadei, R.: Solving the hub location problem in telecommunication network design: a local search approach. *Networks* 44, 94–105 (2004)
6. Rodriguez-Martin, I., Salazar-Gonzalez, J.: Solving a capacitated hub location problem. *European Journal of Operational Research* 184, 468–479 (2008)
7. Ernst, A., Krishnamoorthy, M.: Solution algorithms for the capacitated single allocation hub location problem. *Annals of Operations Research* 86, 141–159 (1999)
8. Randall, M.: Solution approaches for the capacitated single allocation hub location problem using ant colony optimization. *Computational Optimization and Applications* 39, 239–261 (2008)
9. Yaman, H.: Allocation strategies in hub networks. *European Journal of Operational Research* 211, 442–451 (2011)
10. Correia, I., Nickel, S., Saladaha-da-Gama, F.: Hub and spoke network design with single-assignment, capacity decision and balancing requirements. *Applied Mathematical Modelling* 35, 4841–4851 (2011)
11. Dueck, G., Scheuer, T.: Threshold accepting: a general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics* 90, 161–175 (1990)
12. Chen, J.-F.: A hybrid heuristic for the uncapacitated hub location problem. *Omega-The International Journal of Management Science* 35, 211–220 (2007)
13. O’Kelly, M.: A quadratic integer problem for the location of interacting hub facilities. *European Journal of Operational Research* 32, 393–404 (1987)
14. Ernst, A., Krishnamoorthy, M.: Efficient algorithms for the uncapacitated single allocation p-hub median problem. *Location Science* 4, 139–154 (1996)

A Multi-objective Genetic Algorithm Based Handoff Decision Scheme with ABC Supported

Chengbo Zhang, Xingwei Wang, and Min Huang

College of Information Science and Engineering, Northeastern University,
Shenyang, Liaoning 110819, China
wangxw@mail.neu.edu.cn

Abstract. A handoff decision scheme of heterogeneous wireless networks should be comprehensively considered with application quality of service (QoS) requirement, service fee, user preference, mobile terminal condition and access network condition, so as to make users and network providers achieve a win-win situation. With the knowledge of fuzzy mathematics and microeconomics introduced, the influence factors are characterized, and a multi-objective genetic algorithm based handoff decision scheme with always best connected (ABC) supported is proposed. Simulation results show that the proposed scheme is effective with excellent performance.

Keywords: handoff decision, heterogeneous wireless networks, genetic algorithm (GA), always best connected (ABC).

1 Introduction

The Fourth Generation (4G) heterogeneous wireless networks [1] are forthcoming with integrating various wireless access technologies. So wireless networks should enable users to be always best connected (ABC) [2], so as to achieve the global roaming of users in communication networks of different technologies [1].

ABC supports users to connect with the best access network at anytime anywhere according to the user demand. For this 'best' choice, handoff decisions should depend on plenty of factors, such as quality of service (QoS) [3], service fee, and user preference, etc. But QoS requirements are difficult to accurately quantify, and they have very strong fuzziness [4], so that an ability of processing fuzzy information is needed in handoff decision schemes. However, with the commercialization of network nowadays, ABC is not only an unilateral affair of users, but also needs to take user utility and network provider utility into account to realize the maximization of both utilities and avoid frequent handovers, which can lead to the ping-pong effect [5]. In this paper, there are N terminals in M access networks at the same time, and an optimal handoff solution of assigning N terminals to M access networks is found with a lot of factors to make user utility and network provider utility achieve or approach for Pareto optimum solution under Nash equilibrium. Since the number of terminals and access networks will be extraordinarily large in the future application, it is a

multi-objective optimal decision problem with a great deal of computation to select the best one from the M^N handoff schemes, and requires heuristics or intelligent optimization algorithms.

Existing algorithms [6] considered service fee, received signal strength information (RSSI), and user preference, etc. Some algorithms made handoff decision based on total payoff [7], load balancing, and battery lifetime [8], etc. However, most of the works haven't fully considered ABC supporting handoff decision factors. In this paper, application type, application QoS requirement, access network and mobile terminal are characterized with the knowledge of fuzzy mathematics and microeconomics, and they are analyzed attentively. Although it is similar to some other studies [6] that this proposed scheme is regarded as a multiple attribute decision. We try to apply the game analysis to seek a Pareto optimum solution under Nash equilibrium of user utility and network provider utility, and use an elitist selection and individual migration of multi-objective genetic algorithm [9] to improve decision efficiency. So that handoff decision technologies could be widely used.

2 Model Description

2.1 Types of Applications and QoS Requirement

According to the differentiated services (DiffServ) model [10], assume that there are I different types of applications and $ATS = \{AT_1, AT_2, \dots, AT_I\}$ expressing the set of application types. The QoS requirement of each application has four parameters, bandwidth, delay, delay jitter and error rate. And it is denoted by QS_i , where $QS_i = \langle [BW_i^l, BW_i^h], [DL_i^l, DL_i^h], [JT_i^l, JT_i^h], [ER_i^l, ER_i^h] \rangle$ and $i = 1, 2, \dots, I$.

2.2 Access Network Model

Assume that the order number of an access network is $j, 1 \leq j \leq M$, where M denotes the number of access networks.

The access network provider is PI_j and the set of all access network providers is $PIS = \{PI_1, PI_2, \dots, PI_{|PIS|}\}$. Similarly, the access network type and coding scheme are TI_j and CI_j , and the sets of all access network types and all coding schemes are $TIS = \{TI_1, TI_2, \dots, TI_{|TIS|}\}$ and $CIS = \{CI_1, CI_2, \dots, CI_{|CIS|}\}$. CA_j, MV_j, FR_j and TP_j are the coverage, supporting highest velocity, frequency range and lowest signal strength of access network, respectively. The set of application types supported by access network j is $NAS_j \subseteq ATS$. The total bandwidth and available bandwidth of access network are TB_j and AB_j respectively, and the minimum bandwidth threshold of AB_j is AB_j^{\min} . When $AB_j < AB_j^{\min}$, the performance of access network will decline sharply, so it should not accept any of terminal requests to avoid the ping-pong effect.

The set of service levels supported by access network j is SL_j , which can offer many QoS services to various types of applications. And various access networks provide all or part of service levels according to their capacities, so that $SL_j \subseteq SL$.

Assume that it is a k level service that the access network provides to AT_i type of applications, where $k \in SL_j$. The QoS of k level services is $QS_{ji}^k = \langle [bw_{ji}^{kl}, bw_{ji}^{kh}] [dl_{ji}^{kl}, dl_{ji}^{kh}] [jt_{ji}^{kl}, jt_{ji}^{kh}] [er_{ji}^{kl}, er_{ji}^{kh}] \rangle$. The service cost and service fee are ct_{ji}^k and pr_{ji}^k per unit bandwidth and unit time, respectively. For load balancing, pr_{ji}^k is determined by control coefficient and basic fee, which are set by access network providers based on the marketing strategy. Since there are N terminals transferring M access networks, terminals or access networks may cooperate in order to seek greater benefits. If there is any cooperation among terminals or access networks, pr_{ji}^k should be cut or risen.

2.3 Terminal Model

Assume that the order number of a terminal is $t, 1 \leq t \leq N$, where N denotes the number of terminals.

The set of application types and the set of coding schemes supported by terminal t are TAS_t and MCS_t , where $TAS_t \subseteq ATS$ and $MCS_t \subseteq CIS$. WF_t and RS_t are the frequency range of terminal and the lowest signal strength received by terminal t respectively. CV_t is the velocity of terminal at current, and CV_h represents the threshold of high velocity. When $CV_t \geq CV_h$, the terminal is at high velocity condition. Similarly, RC_t and RC_{th} are remainder battery capacity and threshold of low battery capacity. The sequences of user preference in access networks and providers are $PC_t = \{PC_{t1}, PC_{t2}, \dots, PC_{tq}\} \subseteq CIS$ and $PP_t = \{PP_{t1}, PP_{t2}, \dots, PP_{tm}\} \subseteq PIS$, where PC_t and PP_t are sorted by preference degree from high to low. HP_{it} is the service fee that the user is willing to pay for.

In summary, the handoff request of terminal t is $HR_t = \langle AT_i, PC_t, PP_t, HP_{it} \rangle$.

2.4 QoS Satisfaction

There are four coefficients, $\bar{\omega}_i^B, \bar{\omega}_i^D, \bar{\omega}_i^J$ and $\bar{\omega}_i^E$, which represent the importance of bandwidth, delay, delay jitter and error rate respectively, and $\bar{\omega}_i^B + \bar{\omega}_i^D + \bar{\omega}_i^J + \bar{\omega}_i^E = 1$. The tolerance matrix method is utilized to determine the four coefficients. A judgment matrix $A = [a_{mn}]_{4 \times 4}$ is constructed by relative importance between ones of QoS parameters. The rows and columns of A correspond to bandwidth, delay, delay jitter and error rate respectively. The a_{mn} is the approximate value of the parameter weight of m divided by that of n , and $a_{mn} \in (0, 9], a_{mn} = 1 / a_{nm}$, where $m, n = 1, 2, 3, 4$. A tolerance matrix $B = [b_{mn}]_{4 \times 4}$ is

calculated according to A , where $b_{nm} = \sqrt[4]{\prod_{k=1}^4 a_{mk} \cdot a_{kn}}$. The QoS parameter weight ϖ_i^m is calculated as $\varpi_i^m = c_m / \sum_{k=1}^4 c_k$ and $c_m = \sqrt[4]{\prod_{g=1}^4 b_{mg}}$. The ϖ_i^m for AT_i type applications is calculated and assigned to $\varpi_i^B, \varpi_i^D, \varpi_i^J$ and ϖ_i^E , which are held for online. The coefficients are calculated offline only if the importance of QoS parameters has changed. Assume that access network j provides a k level service to terminal t with the AT_i type application running, where k is determined for HR_t by the access network condition and evaluated based on the grey relational analysis method. BS_{ji}^k and DS_{ji}^k are the weighted fuzzy satisfactions for bandwidth and delay, as Equations (1) and (2). Similarly, the weighted fuzzy satisfactions for delay jitter and error rate, JS_{ji}^k and ES_{ji}^k , are calculated only by replacing the relevant parameters of Equation (2) with the corresponding ones.

$$BS_{ji}^k = \varpi_i^B \cdot \frac{\frac{1}{2}(bw_{ji}^{kl} + bw_{ji}^{kh}) - \min_j \frac{1}{2}(bw_{ji}^{kl} + bw_{ji}^{kh})}{\max_j \frac{1}{2}(bw_{ji}^{kl} + bw_{ji}^{kh}) - \min_j \frac{1}{2}(bw_{ji}^{kl} + bw_{ji}^{kh})} \tag{1}$$

$$DS_{ji}^k = \varpi_i^D \cdot \frac{\max_j \frac{1}{2}(dl_{ji}^{kl} + dl_{ji}^{kh}) - \frac{1}{2}(dl_{ji}^{kl} + dl_{ji}^{kh})}{\max_j \frac{1}{2}(dl_{ji}^{kl} + dl_{ji}^{kh}) - \min_j \frac{1}{2}(dl_{ji}^{kl} + dl_{ji}^{kh})} \tag{2}$$

Assume that the ideal service level of AT_i type applications is k^* , where the bandwidth satisfaction is $BS_i^{k^*} = \max_j BS_{ji}^k$. Similarly, $DS_i^{k^*}, JS_i^{k^*}$ and $ES_i^{k^*}$ are obtained. In this paper, the grey relational degree from k level service to k^* level ideal service is regarded as the evaluation QP_{ji}^k of k level service, where $QP_{ji}^k = (GR_{ji}^{kB} + GR_{ji}^{kD} + GR_{ji}^{kJ} + GR_{ji}^{kE}) / 4$ and GR_{ji}^{kB} is calculated as Equation (3). Similarly, $GR_{ji}^{kD}, GR_{ji}^{kJ}$ and GR_{ji}^{kE} are calculated only by replacing the relevant parameters.

$$GR_{ji}^{kB} = \frac{\min_i \min_j |BS_{ji}^k - BS_i^{k^*}| + \frac{1}{2} \min_i \max_j |BS_{ji}^k - BS_i^{k^*}|}{|BS_{ji}^k - BS_i^{k^*}| + \frac{1}{2} \min_i \max_j |BS_{ji}^k - BS_i^{k^*}|} \tag{3}$$

CB_{ji}^k and CD_{ji}^k are the suitability of bandwidth and delay to bandwidth request and delay request of AT_i type applications, as Equations (4) and (5), where β is a constant. Similarly, suitability CJ_{ji}^k and CE_{ji}^k of delay jitter and error rate are calculated only by replacing the relevant parameters of Equation (5). The suitability of QoS for k level service is denoted by CQ_{ji}^k , where $CQ_{ji}^k = \varpi_i^B \cdot CB_{ji}^k + \varpi_i^D \cdot CD_{ji}^k + \varpi_i^J \cdot CJ_{ji}^k + \varpi_i^E \cdot CE_{ji}^k$.

$$CB_{ji}^k = \left(\frac{bw_{ji}^{kl} + bw_{ji}^{kh} - BW_i^l}{2} - \frac{BW_i^h - BW_i^l}{2} \right)^\beta \quad (4)$$

$$CD_{ji}^k = \frac{1}{2} - \frac{1}{2} \sin \frac{\pi}{DL_i^h - DL_i^l} \left(\frac{dl_{ji}^{kh} + dl_{ji}^{kl}}{2} - \frac{DL_i^h + DL_i^l}{2} \right) \quad (5)$$

In summary, the QoS satisfaction is calculated as $SQ_{ji}^k = QP_{ji}^k \cdot CQ_{ji}^k$.

2.5 Other Satisfaction or Suitability

This paper describes the other satisfactions or suitability as follows:

$SP_{i,j}$ is the satisfaction of user for price pr_{ji}^k . And if $pr_{ji}^k \leq HP_{it}$, $SP_{i,j} = 1$; otherwise, $SP_{i,j} = 0$. SC_{ij} is the satisfaction of user for coding scheme CI_{ij} . If $CI_{ij} \in PC_t$, $SC_{ij} = 1/x^2$; otherwise, $SC_{ij} = 0$. Where x is the order number of CI_{ij} in PC_t . SR_{ij} is the satisfaction of user for access network provider PI_{ij} . If $PI_{ij} \in PP_t$, $SR_{ij} = 1/y^2$; otherwise, $SR_{ij} = 0$. Where y is the order number of PI_{ij} in PP_t . SV_{ij} is the velocity suitability. If the velocity of a terminal is high at current, an access network with larger coverage should be considered to transfer for decreasing the number of handovers. It can help to avoid the ping-pong effect. So that if $CV_t < CV_h$, $SV_{ij} = 1$; if $CV_h \leq CV_t \leq MV_j$, $SV_{ij} = 1/z^2$; otherwise, $SV_{ij} = 0$. Where z is the order number of access network type TI_j in TIS sorted by the coverage from high to low. SY_{ij} is the battery capacity suitability. If the battery capacity of a terminal is low at current, an access network with smaller coverage should be considered to transfer for decreasing the power of receiving and sending, and it can help to extend terminal working time. So that if $RC_t \leq RC_{th}$, $SY_{ij} = 1/(|TIS| - z)^2$; otherwise, $SY_{ij} = 1$.

2.6 Game Analysis and Utility Calculations

In this paper, the players are a terminal and an access network. Hence the payoff matrices of terminal and access network are denoted by TG and NG , as Equations (6) and (7). Where the rows or columns of matrix express that terminal t or access network j is willingness and unwillingness to accept in turn. The minus sign shows the lost payoff, and the penalty factor is $\nu > 1$. If a pair of policies $\{a_{i^*}, b_{j^*}\}$ satisfies $TG_{i^*j^*} \geq TG_{i^*j} \wedge NG_{i^*j^*} \geq NG_{ij^*}$, the terminal payoff and access network payoff achieve Nash equilibrium, where $i^*, j^*, i, j = 1, 2$.

$$TG = \begin{bmatrix} HP_{ii} - pr_{ji}^k & HP_{ii} - pr_{ji}^k \\ -v \cdot (HP_{ii} - pr_{ji}^k) & -(HP_{ii} - pr_{ji}^k) \end{bmatrix} \quad (6)$$

$$NG = \begin{bmatrix} pr_{ji}^k - ct_{ji}^k & -v \cdot (pr_{ji}^k - ct_{ji}^k) \\ pr_{ji}^k - ct_{ji}^k & -(pr_{ji}^k - ct_{ji}^k) \end{bmatrix} \quad (7)$$

A coefficient matrix $\Lambda = [\lambda_1 \lambda_2 \cdots \lambda_6]$ is introduced in this paper, $\sum_{i=1}^6 \lambda_i = 1$. The elements of Λ express the relative importance of factors, such as application QoS requirement, service fee, access network coding scheme, access network provider, terminal velocity and terminal battery capacity. The value of λ_i is also determined by the tolerance matrix method, similarly with Section 2.4.

An evaluation matrix $G_{t,j} = [SQ_{ji}^k \ SP_{t,j} \ SC_{t,j} \ SR_{t,j} \ SV_{t,j} \ SY_{t,j}]^T$ and a control coefficient Ω are introduced. If both sides achieve Nash equilibrium, $\Omega = 1$; otherwise, $\Omega = 0$. The user utility and network provider utility are $uu_{t,j}$ and $nu_{t,j}$. If terminal t transfers access network j , $uu_{t,j} = \Omega \cdot \Lambda \cdot G_{t,j} \cdot ((HP_{ii} - pr_{ji}^k) / HP_{ii})$ and $nu_{t,j} = \Omega \cdot \Lambda \cdot G_{t,j} \cdot ((pr_{ji}^k - ct_{ji}^k) / pr_{ji}^k)$; otherwise, $uu_{t,j} = 0$ and $nu_{t,j} = 0$.

2.7 Mathematical Model

The mathematical description of optimal handoff solution is shown as follows.

$$\text{maximize} \left\{ \sum_{t=1}^N \sum_{j=1}^M uu_{t,j} \right\} \quad (8)$$

$$\text{maximize} \left\{ \sum_{t=1}^N \sum_{j=1}^M nu_{t,j} \right\} \quad (9)$$

$$\text{maximize} \left\{ \sum_{t=1}^N \sum_{j=1}^M (uu_{t,j} + nu_{t,j}) \right\} \quad (10)$$

3 Algorithm Design

In this paper, the elitist selection and individual migration of multi-objective genetic algorithm is applied, because it converges more rapidly with keeping diversity.

3.1 Definitions and Operation Rules

S is the size of each population, and each individual $in_q = \langle ch_{q_1}, ch_{q_2}, \dots, ch_{q_N} \rangle$ is a solution of handoff decision problem. Chromosome $ch_{q_t} = \langle AN_{q_t}, k_{q_t} \rangle$ expresses access network AN_{q_t} provides a k_{q_t} level service to terminal t in the q^{th} solution, where $1 \leq q \leq S, 1 \leq t \leq N, 1 \leq AN_{q_t} \leq M, k_{q_t} \in SL$.

The population initialization is a process of assigning an initial value to each individual of populations, where AN_{q_t} and k_{q_t} from each chromosome ch_{q_t} of each

individual in_q are randomly assigned with the values of interval $[1, M]$ and interval $[1, |SL_{AN_{q_i}}|]$ respectively. If $\forall t ((TAS_t \subseteq NAS_{AN_{q_i}}) \wedge (MCS_t \cap CI_{AN_{q_i}} \neq \emptyset) \wedge (WF_t \subseteq FR_{AN_{q_i}}) \wedge (RS_t \leq TP_{AN_{q_i}}) \wedge (CV_t \leq MV_{AN_{q_i}}) \wedge (HP_{ti} \geq pr_{AN_{q_i}}^k) \wedge ((AB_{AN_{q_i}} - bw_{AN_{q_i}}^{kh}) \geq AB_{AN_{q_i}}^{\min}))$, handoff solution in_q is feasible. $FT_U(in_q)$, $FT_N(in_q)$ and $FT_{UN}(in_q)$ are fitness functions for Equations (8)-(10). If in_q is feasible, $FT_U(in_q) = \sum_{i=1}^N (1/uu_{i,AN_{q_i}})$, $FT_N(in_q) = \sum_{i=1}^N (1/nu_{i,AN_{q_i}})$, and $FT_{UN}(in_q) = \sum_{i=1}^N (1/uu_{i,AN_{q_i}} + 1/nu_{i,AN_{q_i}})$; otherwise, $FT_U(in_q) = FT_N(in_q) = FT_{UN}(in_q) = +\infty$.

In this paper, a linear crossover is utilized and a crossover probability is CP , where $0 < CP < 1$. Assume that two parent individuals are in_{q_1} and in_{q_2} taking part in the crossover, and ρ is a random pure decimal. If $\rho > CP$, copy the parent individuals as child individuals. Otherwise, execute the linear crossover operation and generate two child individuals in'_{q_1} and in'_{q_2} . The linear crossover operation of chromosome is that $AN'_{q_1} = (\alpha_1 AN_{q_1} + (1 - \alpha_1) AN_{q_2})$, $k'_{q_1} = (\alpha_1 k_{q_1} + (1 - \alpha_1) k_{q_2})$, $AN'_{q_2} = (\alpha_2 AN_{q_1} + (1 - \alpha_2) AN_{q_2})$, and $k'_{q_2} = (\alpha_2 k_{q_1} + (1 - \alpha_2) k_{q_2})$, where $0 \leq \alpha_1, \alpha_2 \leq 1$.

The same mutation probability occurs in every chromosome of individuals, and it is denoted by MP , where $0 < MP < 1$. Randomly select a chromosome ch_{q_i} from an individual in_q , and then randomly generate a pure decimal δ . If $\delta > MP$, not execute the mutation operation. Otherwise, assign random values which are in the domains to AN_{q_i} and k_{q_i} of chromosome ch_{q_i} respectively.

3.2 Algorithm Description

Step 1: Set the crossover probability is CP , the mutation probability is MP , the size of the tournament race is RS , the size of subpopulations is S , the upper limit of the individual copy number is RT , and the time of continuous iterations is IT .

Step 2: According to Section 3.1, generate the subpopulation P_U , P_N and P_{UN} for the three optimal objectives of Equations (8)-(10). The elitist populations of them are $ES_U = ES_N = ES_{UN} = \emptyset$, the minimum fitness of subpopulations are $FT_U^* = FT_N^* = FT_{UN}^* = +\infty$, and the global minimum fitness is $FT^* = \min\{FT_U^*, FT_N^*, FT_{UN}^*\} = +\infty$.

Step 3: Let $it = 1$.

Step 4: According to Section 2.6, make the terminal and access network from each chromosome of the individuals of P_U , P_N and P_{UN} play the game. And then determine the utilities of both sides according to whether to accept cooperation.

Step 5: Calculate the fitness of each individual of P_U, P_N and P_{UN} , and denote the minimum fitness of each population by $P_{U_{min}}, P_{N_{min}}$ and $P_{UN_{min}}$. If $\min\{P_{U_{min}}, P_{N_{min}}, P_{UN_{min}}\} < FT^*$, $FT^* = \min\{P_{U_{min}}, P_{N_{min}}, P_{UN_{min}}\}$ and $it = 1$. If $P_{U_{min}} < FT_U^*$, use all the corresponding individuals of $P_{U_{min}}$ to displace ES_U and $FT_U^* = P_{U_{min}}$; if $P_{U_{min}} = FT_U^*$, unite ES_U with all the corresponding individuals of $P_{U_{min}}$; execute the same operation to ES_N and ES_{UN} , and then $ES = ES_U \cup ES_N \cup ES_{UN}$.

Step 6: If $it > IT$, go to Step 10.

Step 7: Initialize the copy number of each individual with 0 in P_U, P_N, P_{UN} and ES . Randomly select RS individuals from P_U and ES with ensuring that all the copy numbers of the individuals selected are less than RT , select the individual with minimum fitness as a parent individual of P_U and the copy number of it is increased by 1. Execute the same operation to P_N and P_{UN} . Execute repeatedly the operation until parent population P_U, P_N and P_{UN} are generated.

Step 8: Execute the crossover and mutation to P_U, P_N and P_{UN} respectively, so as to generate the new generation subpopulation P_U, P_N and P_{UN} .

Step 9: Let $it = it + 1$, go to Step 4.

Step 10: If $FT^* \neq +\infty$, regard the individual of ES corresponding to FT^* as the handoff decision solution, successfully finish; otherwise, there is no solution, failed.

4 Simulation Results and Discussions

In the network simulator 2 (NS2), we simulate the scheme of this paper (Scheme I), the greedy algorithm based handoff decision scheme [11] (Scheme II) and the QoS based handoff decision scheme [12] (Scheme III) in three hexagonal honeycomb topologies, which are covered by many different types of access networks. We use 3, 5, 10, 20 and 50 terminals to execute 500 times handoff decisions respectively, and set two user cases in each different number of terminals, according to Table 1.

Table 1. User cases

User cases	Coding scheme (%)			Network provider (%)			Velocity (%)		Battery capacity (%)		
	PC ₁	PC ₂	PC ₃	PP ₁	PP ₂	PP ₃	high	medium low	high	medium	low
1	33	33	33	33	33	33	33	33	33	33	33
2	60	20	20	60	20	20	60	20	20	20	60

We make comparisons and regard the means as simulation results in Fig. 1. As shown in Fig. 1a, Scheme III has the best performance in QoS satisfaction, but Scheme I is still better than Scheme II. Because Scheme I gives consideration to plenty of factors and QoS is just one of them. Scheme III only seeks the QoS satisfaction of user. Fig. 1b-1h shows Scheme I has better performances than the others in these factors, such as user utility and network provider utility, Pareto optimum solution ratio, price satisfaction, coding scheme satisfaction, access network provider satisfaction, velocity suitability, and battery capacity suitability. As shown in Fig. 1i, Scheme I takes a bit of time acceptable. And the larger the number of handoff requests is, the more rapid this genetic algorithm is than the greedy algorithm.

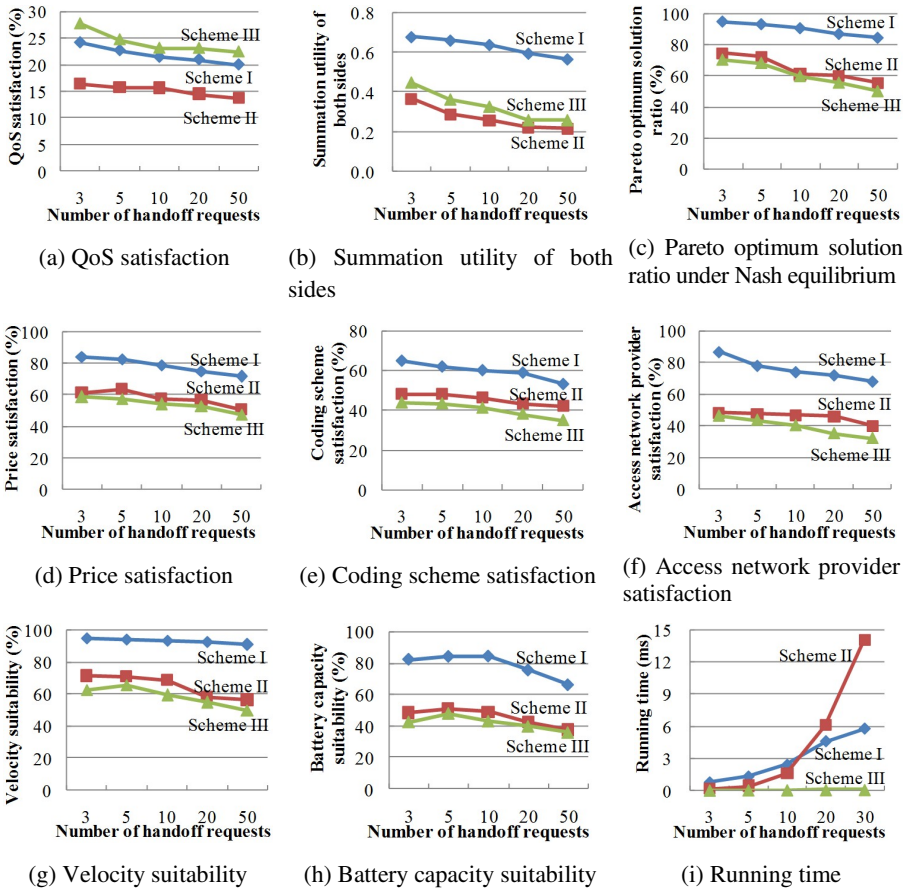


Fig. 1. Performance analyses among Scheme I , II and III

5 Conclusions

In this paper, the handoff decisive factors are synthetically considered, and a multi-objective genetic algorithm based handoff decision scheme with ABC supported is

proposed. And the simulation results show that the proposed scheme has better performance compared with existing schemes in most of the quotas.

Acknowledgements. This work is supported by the National Science Foundation for Distinguished Young Scholars of China under Grant No. 61225012; the National Natural Science Foundation of China under Grant No. 61070162, No. 71071028 and No. 70931001; the Specialized Research Fund of the Doctoral Program of Higher Education for the Priority Development Areas under Grant No. 20120042130003; the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20100042110025 and No. 20110042110024; the Specialized Development Fund for the Internet of Things from the ministry of industry and information technology of the P.R. China; the Fundamental Research Funds for the Central Universities under Grant No. N110204003 and No. N120104001.

References

1. Yan, X.H., Şekercioglu, Y.A., Narayanan, S.: A Survey of Vertical Handover Decision Algorithms in Fourth Generation Heterogeneous Wireless Networks. *Computer Networks* 54(11), 1848–1863 (2010)
2. Gazis, V., Alonistioti, N., Merakos, L.: Toward a Generic “Always Best Connected” Capability in Integrated WLAN/UMTS Cellular Mobile Networks (and Beyond). *IEEE Wireless Communications* 12(3), 20–29 (2005)
3. Song, J., Chang, M.Y., Lee, S.S., Joung, J.: Overview of ITU-T NGN QoS Control. *IEEE Communications Magazine* 45(9), 116–123 (2007)
4. Wang, X.W., Hou, M.J., Huang, M., Wang, Q.: A Game Theory Based Fuzzy Unicast QoS Routing Scheme. *Chinese Journal of Computers* 30(1), 10–17 (2007) (in Chinese)
5. Chang, B.J., Chen, J.F.: Cross-layer-based Adaptive Vertical Handoff with Predictive RSS in Heterogeneous Wireless Networks. *IEEE Transactions on Vehicular Technology* 57(6), 3679–3692 (2008)
6. Çalhan, A., Çeken, C.: Case Study on Handoff Strategies for Wireless Overlay Networks. *Computer Standards and Interfaces* 35(1), 170–178 (2013)
7. Liu, X.W., Fang, X.M., Chen, X., Peng, X.S.: A Bidding Model and Cooperative Game-based Vertical Handoff Decision Algorithm. *Journal of Network and Computer Applications* 34(4), 1263–1271 (2011)
8. Zhu, S.F., Liu, F., Chai, Z.Y., Qin, Y.T., Wu, J.S.: Simple Harmonic Oscillator Immune Optimization Algorithm for Solving Vertical Handoff Decision Problem in Heterogeneous Wireless Network. *Acta Physica Sinica* 61(9), 096401, 1–10 (2012) (in Chinese)
9. Qi, R.B., Qian, F., Du, W.L., Yan, X.F.: Multiobjective Genetic Algorithm Based on Elitist Selection and Individual Migration. *Control and Decision* 22(2), 164–168 (2007) (in Chinese)
10. Lin, C., Shan, Z.G., Sheng, L.J., Wu, J.P.: Differentiated Services in the Internet: a Survey. *Chinese Journal of Computers* 23(4), 419–433 (2000) (in Chinese)
11. Hu, Y.J., Zhu, J.K.: A Multiuser Detector Based on Greedy Algorithm for CDMA Communication. *Computer Engineering and Applications* 39(17), 17–19 (2003) (in Chinese)
12. Song, Q.Y., Jamalipour, A.: A Network Selection Mechanism for Next Generation Networks. In: *Republic of IEEE International Conference on Communications*, vol. 2, pp. 1418–1422 (2005)

A Weight-Based Graph Coloring Approach to Airport Gate Assignment Problem

YanJun Jiang and Xueyan Song

School of Computer Science & Technology, Tianjin University, Tianjin 300072

Abstract. A graph coloring model of airport gate assignment problem (AGAP) is constructed in this paper, and a kind of new time slot algorithm is used to find out the conflict sets of all the flights. By considering both the type and time conflict of a flight comprehensively, a new criterion for measuring the difficulty of a flight is put forward which can be used to compute the difficulty coefficients of each flight to be assigned. And after that all these flights will be sorted in descending order by their difficulty coefficients before assigning them to available gates. Finally, experimental results on the stochastic data sets demonstrate the effectiveness of the algorithm put forward in this paper.

Keywords: Gate Assignment, Vertex Coloring, Difficulty Coefficients.

1 Introduction

With the rapid development of society and economy, there is a growing demand for air travelling, which greatly enlarges the scale of civil aviation and results in an increasing number of flights. However, the construction of airport gates is far lagged behind. Therefore, the contradiction between the growing number of flights and the inadequacy of gates becomes more and more severe. Under the circumstances, we need a more excellent algorithm to address the optimization problem of assigning fast increasing number of flights to limited number of airport gates. And numerous scholars have done great amount of research work in this field. Oversea Researchers have a profound understanding of this problem and have obtained a multitude of outstanding optimization results. To sum up, their researches come down to three categories: Mathematical Programming Approach [1-2], Expert System [3-4] and System Simulation [5-6]. In China, scholars began doing research in this filed much later but we have also achieved some excellent research results. For instance, Xueming Zhang and Fazhong Shi [7] put forward and implemented Three-Level-Reasoning -Model Airport Gate Automatic Expert System; Chen Tian, Guixi Nai [8] utilized the Genetic Algorithm to solve the AGAP; Jun Wen, Bing Li[9] etc. introduced the time slot division approach to define the clash sets of flights and they also converted the AGAP to graph coloring problem; Rongwu Luo, Ruhe Xie[10] did further research on the results of Wen & Xie, and they put forward an improved time slot algorithm and use a new division method and coloring algorithm to optimize the assignment results. In this paper, we do further research on the basis of Wen and Luo's research results and we propose an innovate time slot division approach and bring in the new concept of Difficulty Coefficients, which produce better results of AGAP than theirs.

2 The Graph Coloring Model for AGAP

2.1 AGAP

According to the size of flights, we can divide the flights into three categories: large size, medium size and small size. Similarly, we can separate the gates in the airport to three categories: the large type, the middle type and the small type. The AGAP means that assigning appropriate gates for all the arriving and leaving flights at an airport to ensure all the flights run on time. This issue belongs to constraint optimization problems and it is NP-Complete, thus no simple polynomial time algorithm for it.

2.2 Constraints for AGAP

The constraints that need to be satisfied when assigning a gate are as follows:

1. There can be at most one flight on one gate at the same time interval.
2. Every flight must be assigned to one gate, and also at most one gate.
3. The minimum grounding time constraint must be satisfied. That is, the time interval between the flight arriving and leaving should be longer than the minimum ground time.
4. The successive flights at the same gate should keep a safe time interval to secure flights entering and leaving smoothly (this is a soft constraint).
5. The flights and gates mutual matching constraint should be satisfied. In other words, the large flights can only be assigned to large gates, middle flights can be assigned to both large and middle gates, and small flights can be assigned to any type of gates.

2.3 Graph Coloring Model for AGAP

The AGAP can be converted into an undirected graph coloring optimization problem [11]. In this problem, the flights are viewed as vertexes in the graph and gates are considered as the available colors. If there is time conflict between two flights then there will be an edge between the two vertexes in the graph. The optimization problem in this paper is to use the fewest gates to park all the flights. In other words, our aim is to find the coloring number of the undirected graph. This graph coloring model can be illustrated in the following formal language. $G(V, E)$ where, $V = \{v_i \mid \text{all the flights need to be assigned } i\}$, $E = \{(v_i, v_j) \mid \text{if there is time conflict between vertexes } v_i, v_j\}$

3 Algorithm for AGAP

The objective function in this paper is to minimize the number of used gates under the condition of that all the flights can be assigned a gate. If the number of gates used is the same, we would consider the one use smaller gates to be better. Since the number of gates in an airport is certain, thus this minimum problem can be transformed into a maximum problem — maximizing the remaining gates number. We

make the following assumption. The number of remaining large, medium and small type gates are denoted as x 、 y 、 z . As it is better to leave behind large type of gates than medium and small type gates, therefore, in order to measure the assignment results more precisely, we give different priority to different type of gates. There are many different priority assignment methods, in this paper, we use the simplest approach. We set the priority of large, medium and small type gates to 3, 2 and 1 separately. The linearity-weighted-sum of remaining gates is utilized as the objective function which is formulated as follows:

$$\text{maximize Score}(x, y, z) = x + 2y + 3z \tag{1}$$

The AGAP assignment algorithm is divided into three steps:

1. Define the time conflicts set between all the flights.
6. Sort all the flights according to their difficulty coefficients.
7. Color the vertexes in the undirected graph.

3.1 Algorithm for Defining the Time Conflicts Set between All the Flights

Inspired by the method put forwarded in the paper [10], we designed an innovative approach for defining the conflict time set. The new method does not need to sort all the flights according to their arriving and departing time. So it is easy to be implemented. Assuming that all the flights set need to be assigned is $V = \{v_1, v_2, \dots, v_N\}$, here N is an integer which signifies the number of flights to be assigned. b_i is used to indicate the beginning time of flight i , and all the b_i compose the beginning time set $B = \{b_i | i = 1, 2, \dots, N\}$. Similarly, the finishing time of flights i is denoted by f_i , and the finishing time set $F = \{f_i | i = 1, 2, \dots, N\}$ is comprised of all the f_i . All the flights that conflict with flight i ($j > i$) is signified as $S_i (i = 1, 2, \dots, N)$, $S = S_1 \cup S_2 \cup \dots \cup S_N$.

Algorithm 1. Pseudo-code

```

Input: B, E
for = i to i <= N
  for = j to j <= N
    if  $b_i == b_j$ 
      add  $v_j$  to  $S_i$ 
    else if  $b_i < b_j$ 
      if  $f_i > b_j$ 
        add  $v_j$  to  $S_i$ 
      else if  $b_i > b_j$ 
        if  $b_i < f_j$ 
          add  $v_j$  to  $S_i$ 
    end for
  end for
Output: S
    
```

Fig. 1. Algorithm for defining the time conflicts set

The defining method for conflicts between flight time is as follows: if two flight i and flight j arrive at the same time then they conflict with each other thus they cannot be scheduled to the same gate at the same time; if the arriving time of flight i is between the arriving time and departing time of flight j then they also conflict; if the arriving time of flight j is between the arriving time and departing time of flight i which also means that there is conflict between them.

The pseudo code of this algorithm is listed in Fig.1

3.2 Sorting Algorithm of Flights Difficulty Coefficients

In this paper, we would assign the flights according to their difficulty coefficients, which depend on two aspects: the number of conflicts of the flight with others and the size of this flight. Namely, large size flights are more difficult to assign than medium size flights and medium size flights are more difficult to assign than small size flights. In paper [9] the author assign flights according to their conflicts number, if two flights have the same conflicts number then the size of flights will be considered. This means that conflicts number is more important than flights size. Conversely, paper [10] gives more priority to size of flights, they first schedule flights according to their size, if two flights are the same size then they will consider their conflicts number. However, according to our experiments and analysis, we believe that both two methods have advantages and disadvantages. They adapt to different flights data set. By considering both sides comprehensively, we put forward an innovative method for measuring difficulty of assigning flights. The difficult coefficients are formulated as follows:

$$difficulty_i = \alpha \frac{type_i}{max_type} + (1 - \alpha) \frac{degree_i}{max_degree} \quad (2)$$

where:

$difficulty_i$: The difficulty of flight i ;

$\alpha(0 \leq \alpha \leq 1)$: the weight allocated to the size of flights;

$type_i$: The type flight i , and:

$$type_i = \begin{cases} 1 & \text{flight } i \text{ is small type} \\ 2 & \text{flight } i \text{ is medium type} \\ 3 & \text{flight } i \text{ is large type} \end{cases}$$

max_type : Value of largest flight;

$degree_i$: Conflicts number of flight i with other flights;

max_degree : The maximum of time conflicts number for all flights.

We can assign different weight to the conflicts number and type of flights by adjusting the value. Therefore, we can change the priority for assigning all the flights.

From the formulation of definition (2), we know that the less the value, the less important of flights type but more important of conflicts number, which is suitable for flights with more conflicts number and smaller size but adverse to flights with less conflicts and larger size, and vice versa. By changing value dynamically, we can find a much better results for different flights set. The pseudo code for sorting algorithm of flights difficulty coefficients is listed below.

```

Algorithm 2. Pseudo - code
Calculate the degree of each flight
for  $k = 1$  to  $k \leq N$ 
  for  $i = 1$  to  $i \leq N$ 
    if  $degree_i > \max\_degree$ 
       $\max\_degree = degree_i$ 
    end for
  for  $i = 1$  to  $i \leq N$ 
     $difficulty_i = \alpha * type_i +$ 
       $(1 - \alpha) * degree_i / \max\_degree$ 
  end for
  sort all the remaining vertexes according
  to their difficulty coefficients
  assign the sequence number  $k$  to the flight
  with maximum difficulty coefficients
  subtract one to the degrees of all its adjacent vertexes
  recalculate the degree of each flight
end for

```

Fig. 2. Graph 2, sorting algorithm of flights difficulty coefficients

3.3 Graph Coloring Algorithm of AGAP

We can calculate all the sequence number using the method put forward in 3.2, and then we will sort all the flights according to their sequence number and then we obtain the assigning order for all flights, from 3.2 we know that the order is the descending order of flights assignment difficulty coefficients. Thereafter, the algorithm proposed in paper [10] is utilized to color all the flights vertexes. So as to search the best assignment, different α value is used to adjust the weight between the two aspects of flights assignment. Therefore, the best result can be chosen during the variation of α .

4 Experimental Results

4.1 Experimental Data Set Generation

In order to validate the new algorithm that we put forward, ten different data sets are collected to compare the results among paper [9], [10] and ours. The flights data sets

are randomly generated by computer program. And each data set is composed of flights data and gates data.

1. **Flights data sets generation method.** By setting 5 minutes as the time unit for flights, the arriving times of flights are produced firstly, which follow uniform distribution between interval $[0, 48]$. This can be depicted by formal language, that is, assuming X is an integral random variable, obeying the uniform distribution between interval $[0, 48]$, a value for X is generated randomly at first and then $5 \cdot X$ is used as the arriving time for the flight. Similarly, the parking time for flights can be obtained. In our experiment, we assume that the flights parking time is between 30 minutes and 90 minutes. So if Y is denoted the integral random variable for parking time which follow the uniform distribution between interval $[6, 18]$, a value for Y is generated randomly at first then $5 \cdot Y$ is used as the parking time for the flight. Finally, the departing time of the flight can be acquired by adding the parking time to beginning time.
8. **Gate data generation method.** Each flight has two attributes: the gate number and the gate type. The former can be obtained by giving number to the flights according to their generation sequence. The gate type can be gotten by randomly choosing from 1, 2 and 3.

4.2 Experimental Results

The symbol "--" in the above table signifies no solution can be found. The best results in the table are in bold typeface. From the table, we can conclude that algorithms in paper [9] and [10] have its own advantage, but our method is better than both of theirs. Therefore, we know from the experimental results that the best choice for the AGAP is not to give too much priority to each aspects of the problem, but to consider them comprehensively and select the best results through adjusting the weight factor.

Table 1. Experimental Results

Data set	Flight num	Gate num	[9]	[10]	ours
Data 1	50	25	5	--	6
Data 2	60	30	21	20	23
Data 3	80	40	29	38	39
Data 4	100	50	24	31	31
Data 5	100	50	--	--	34
Data 6	100	60	46	55	58
Data 7	100	70	77	78	80
Data 8	120	60	19	28	28
Data 9	200	100	68	79	84
Data 10	200	100	81	81	85

The results are based on the data sets randomly generated using the method in 4.1.

The symbol "--" in the above table signifies no solution can be found. The best results in the table are in bold typeface. From the table, we can conclude that algorithms in paper [9] and [10] have its own advantage, but our method is better than both of theirs. Therefore, we know from the experimental results that the best choice for the AGAP is not to give too much priority to each aspects of the problem, but to consider them comprehensively and select the best results through adjusting the weight factor.

5 Conclusion

In this paper, we conduct further research on the forefathers in this field and establish a graph coloring method for AGAP. What's more, a new comprehensive method of measuring flights assignment difficulty is initially used. By adjusting the α value, we can find a most suitable value for each data set so as to minimize the gates number occupied. And the experimental results confirm that our algorithm is superior than the method used in [9],[10], which means our method can boost the flights utilization rate and reduce the operating fees of airport.

Acknowledgement. This study is supported by grants from National Natural Science Foundation of China and Civil Aviation Administration of China(61039001) and Tianjin Municipal Science and Technology Commission(11ZCKFGX04200).

References

1. Xu, J.F., Bailey, G.: The Airport Gate Assignment Problem: Mathematical Model and a Tabu Search Algorithm. In: Proceedings of the 37th Hawaii International Conference on System Sciences (2001)
2. Ding, H., Lim, A., Rodrigues, B.: Aircraft and Gate Scheduling Optimization at Airports. In: Proceedings of the 37th Hawaii International Conference on System Sciences (2004)
3. Geun-Sik, J., Jong, J.J., Chang, Y.Y.: Expert System for Scheduling an Airline Gate Allocation. *Expert Systems with Application* 13(4), 275–282 (1997)
4. Soi, H.L., Jia, M.C., Henry, F.: Development of an Intelligent Agent for Airport Gate Assignment. *Journal of Air Transportation* 7(2), 103–114 (2002)
5. Cheng, Y.: A Ruled-Based Reactive Model for the simulation of Aircraft on Airport Gate. *Knowledge-Based Systems* 10(4), 225–236 (1998)
6. Chang, Y.: A Network Model for Gate Assignment. *Journal of Advanced Transportation* 32(2), 176–189 (1998)
7. Xue, M.Z., Shi, F.Z.: Research on Airport Gate Assignment Expert System. *Computer Engineering* (6), 69–71 (2000)
8. Tian, C., Nai, G.X.: Genetic-based Airport Gate Assignment Strategy. *Computer Engineering* 31(3), 186–188 (2005)
9. Wen, J., Li, B., Wang, Q.R., Du, W.: Airport Gate Assignment Graph Coloring Model and Algorithm. *Systems Engineering Theory Methodology and Applications* 14(2), 136–140 (2005)
10. Luo, R.W., Xie, R.H., Zhang, D.Z.: Airport Gate Assignment Graph Coloring Model and Algorithm. *System Engineering Theory and Practice* (November 2007)

Game-Based Scheduling Algorithm to Achieve Optimize Profit in MapReduce Environment

Cong Wan, Cuirong Wang, Ying Yuan, and Haiming Wang

College of Information Science and Engineering,
Northeastern University, Shenyang 110044, China
10000Cong@163.com

Abstract. MapReduce is a programming model and an associated implementation for processing and generating large data sets. Providing MapReduce as a service is the development future trend. By leveraging the game theory, this paper proposes a scheduling algorithm to deal with the competition for resources between multiple jobs in MapReduce. Firstly, we present a model that could estimate job executing time, and then a utility function of job and an optimization objective are brought forward; thirdly, we present a game model to solve the optimization problem. The proof and the solution are also present. Finally, we implement the algorithm and experiment it in a hadoop cluster. The result shows the present algorithm could schedule jobs rational.

Keywords: scheduling, MapReduce, game model, QoS.

1 Introduction

Cloud computing is composed of a series of technology, such as programming model, virtual machine monitor, topology and security. MapReduce[1] is a widely used programming model and an associated implementation for processing and generating large data sets. It divided a job into map and reduce stages which corresponding to the map task and reduce task. Files were divided into 64MB blocks, and several copies of each block were stored on different machines. One of the core ideas of MapReduce is to process data locally in order to save network bandwidth, so each map task corresponding to a block and MapReduce master attempts to schedule a map task on the machine that contains a replica of the block. A global Synchronization will transport the output of the map task to the reduce task. Finally, reduce task will output the final result.

MapReduce split job into many tasks, and scheduler in MapReduce determines which task should be executed on a certain node in the cluster. To guarantee a good quality of service (QoS) in the framework a good scheduler should be used, so as not to cause important or urgent jobs can not get the resources. The two important measurements for QoS is time and cost. Users' satisfaction is decisive to them together. Users have different preference. Some of them are willing to give up response time in exchange for low cost, on the other hand, some of them are contrary.

In this paper, we present a multi-job scheduling algorithm in MapReduce based on game theory. It deals with the problem of competition for slot between jobs. We present a model to estimates the executing time of job given the price, and then we set up a slot allocation optimal model. Finally, we solution the model using game theory and develop a scheduler in MapReduce environment.

2 Relate Works

Game theory is widely used in resource allocation and job scheduling. LI Ming Chu et al. [3] propose a grid resource auction model which is multi-winners and based on the uncompleted information game theory. Wang Cong et al. [4] propose a non-cooperative model to deal with the competition for physical bandwidth between virtual networks in virtual data center by leveraging the game theory.

There are some schedulers released with hadoop. FIFO is the default scheduler in hadoop. It select job according to the priority of the job and then the arrival time. Capability scheduler [5] supports multi queues which are allocated a fraction of the total resource capability. Fair scheduler [6] allows jobs to obtain the resource fairly with the passage of time.

Many researchers present there schedulers. Thomas Sandholm and Kevin Lai [7] present the Dynamic Priority (DP) parallel task scheduler for hadoop. It allows users to control their allocated capacity by adjusting their spending over time. But there work was not related to user satisfaction. Xicheng Dong et al. [8] present a mixed real-time and non-real-time job scheduler in hadoop. They try to meet users' QoS demands, but only consider the time factor. Jord`a Polo et al. [9] present a model to estimates individual job completion times given a particular resource allocation, and uses these estimates as to meeting job's performance goal. Kamal Kc and Kemafor Anyanwu [10] present a scheduler to meet the deadlines of jobs in hadoop.

3 Research Motivations

With the development of cloud computing, its more and more popular to obtain service from a large data center instead of allocate an expensive server for every individual department. Jobs in MapReduce are shifting from batch processing to multi-user and multi-job. The attendant problem is to guarantee the QoS of user. In this paper we consider job executing time and cost as the measurements for QoS.

In MapReduce framework, scheduler is an important component. Jobs in MapReduce have been divided into map phase and reduce phase, so there are maptask and reducetask two types of task. Both tasks are scheduled by the scheduler, but they have their own characteristics. Each node in the cluster has slots which are assigned to task as resources. As task, there are two types of slots. In this paper, we only consider scheduling in map phase because we assume that reduce task always could get a slot.

4 Model for the System

In this paper, we have the following assumptions: (1)all the computer in the cluster have the same the cluster is homogeneous computing capability.(2) there are always enough reduce slot. (3) Each reduce task process the same amount of data.

Let M represents a collection of map slots. Each running job is assigned a subset of M , s_i is the number of map slots those assigned to job i .

$U_i (s_i)$ is utility function of job i as the profit of the job with s_i map slot. Thus, the map slot allocation Optimization model could be described as equations (1).

$$\begin{aligned}
 &MAX \sum_{i=1}^n U_i (s_i) \\
 &s.t. \forall i, \quad s_i \geq 0, \\
 &\quad \sum_{i=1}^n s_i \leq |M|
 \end{aligned} \tag{1}$$

The sum of s_i is less then $|M|$ and each s_i is greater than 1.

Let $J = \{j_i \mid i = 1, 2, \dots, n\}$ denotes the jobs that submitted to the cluster. F represents of utility function of jobs.

j_i represent a job which is a application based MapReduce, it was developed by users and submitted to the cluster, it related to the following parameters:

AT_i : job arriving time.

MW_i : map workload. We define the workload of MapReduce WordCount programme process a map task which has 64MB data as 1 unit.

MN_i : number of map task.

α_i : the weight of time.

β_i : the weight of cost.

$FinMN_i$: the number of completed map.

P_i : the price of unit time.

smT_i : single map time.

$rate_i$: the proportion of resources should be allocated to Job_i . It can be calculated as equations (2).

$$rate_i = \frac{P_i}{\sum_{i=1}^n P_i} \tag{2}$$

$am_i = rate_i * SlotM$: assigned map slot number.

T_i : running time.

U_0 represents the maximum satisfaction which is defined as a coefficient l multiplied by the workload of the job.

We define U_i as utility value of j_i described as equations (3) to describe the users' satisfaction.

$$U_i = U_0 - \alpha * T - \beta * Cost \quad . \quad (3)$$

The cost of user only considers the cost of the map phase. In the map phase, each maptask must be paid and we assume each maptask spend the same time. The equation (4) describes the cost calculation method.

$$Cost_i = smT_i * MN_i * P_i \quad . \quad (4)$$

There are some parameters about cluster. $CurrT$ represent the current time. $SlotM$ represent the number of map slot in the cluster.

Our algorithm is a share scheduling algorithm which allocated to each Job a certain percentage of the resources and the proportion of resources is decided by the bid. The time of map stage time can be determined by the number of map task and allocated resources. It can be calculated using the equation (5).

$$\begin{aligned} mT_i &= CurrT - AT_i + smT_i * \frac{MN_i - FinMN_i}{am_i} \\ &= CurrT - AT_i + smT_i * \frac{MN_i}{\frac{P_i}{\sum_{i=1}^n P_i} * SlotM} \quad . \quad (5) \\ &= CurrT - AT_i + \frac{smT_i * (MN_i - FinMN_i)}{SlotM} * \frac{\sum_{i=1}^n P_i}{P_i} \end{aligned}$$

In this paper, we only consider about map task, so the job running time can by calculated as equation (6).

$$T_i = mT_i = CurrT - AT_i + \frac{smT_i * (MN_i - FinMN_i)}{SlotM} * \frac{\sum_{i=1}^n P_i}{P_i} \quad . \quad (6)$$

We can use the equation (7) to calculate the utility of the job.

$$\begin{aligned}
 U_i &= U_0 - \alpha_i * T_i - \beta_i * Cost_i \\
 &= MW_i * l - \alpha_i * (CurrT - AT_i + \frac{smT_i * (MN_i - FinMN_i) * \sum_{i=1}^n P_i}{SlotM}) - \beta_i * smT_i * MN_i * P_i
 \end{aligned}
 \tag{7}$$

For simplicity, we define three coefficients for each job which are described as equation (8) (9) (10).

$$a_i = - \frac{\alpha_i * smT_i * (MN_i - FinMN_i)}{SlotM}
 \tag{8}$$

$$b_i = -\beta_i * smT_i * MN_i
 \tag{9}$$

$$c_i = MW_i * l - \alpha_i * (CurrT - AT_i)
 \tag{10}$$

Finally, the utility is defined as equation (11)

$$U_i = a_i * \frac{\sum_{i=1}^n P_i}{P_i} + b_i * P_i + c_i
 \tag{11}$$

So, we can define utility function of j_i as equation (12) depends on equation (11)

$$F_i(p_i, p_{-i}) = a_i * \frac{\sum_{i=1}^n P_i}{P_i} + b_i * P_i + c_i
 \tag{12}$$

$p_{-i} = (p_1 \dots p_{i-1}, p_{i+1} \dots p_n)$ represents the price of other jobs, it would affect the utility function of j_i .

5 The Proof

There is a allocation scheme $\bar{s} = (s_1, s_2, \dots, s_n)$ which $s_i = M * rate_i$ is the optimal solution of equation (1). The allocation scheme is at the Nash equilibrium point of a non-cooperative game. The following is the proof.

For equation (13), there is a Nash equilibrium point only if there is a price scheme p^* meet the condition $F_i(p^*, p_{-i}^*) \geq F_i(\bar{p}, p_{-i}^*)$, \bar{p} is the group of all other possible price of job i.

Let $W_{-i} = \sum_{k \neq i} p_k$, the derivative of equation (12) is equation (13)

$$F_i' = - \frac{a_i * w_{-i}}{p_i^2} + b
 \tag{13}$$

It is strictly monotone decreasing in the intervals of $p_i > 0$. It could be conclusion that $F_i'' > 0$, so there is a unique p^* making $F_i' = 0$ and there is a unique p_i making equation (11) a maximum. By this time allocation scheme \bar{s} is the optimal solution.

6 Performance Evaluation

We set up Hadoop on a computer cluster which has 1 NameNode and 4 DataNodes. Both NameNode and 4 DataNodes are composed of 2.93GHz Intel Core2 Duo CPU and 2GB memory, All nodes were connected by a gigabit switch. In terms of software, we implement the application using jdk-6u26-windows-i586 and eclipse-SDK-3.6.1. The operating system is Ubuntu10.04 and the Hadoop version is Hadoop-1.0.3.

We configure 4 map slot and 2 reduce slot in each tasktracker. Genetic algorithm settings: Crossover probability is 0.35, mutation probability is 0.05, iteration time is 5000 and initial population size is 10.

We executed 3 real job to experiment out scheduling algorithm. The parameters of the three jobs are shown in Table 1. We defined workload of one map task in wordcount as workload unit.

Table 1. Information of jobs

Job name	Job type	Map task number	Map work-load	l	α	β
Job1	wordcount	58	58	5	0.1	0.9
Job2	secondarySort	59	29	5	0.9	0.1
Job3	TeraSort	58	43.07	5	0.5	0.5

Fig. 1 shows the number of slot allocated to the three job changes over time.

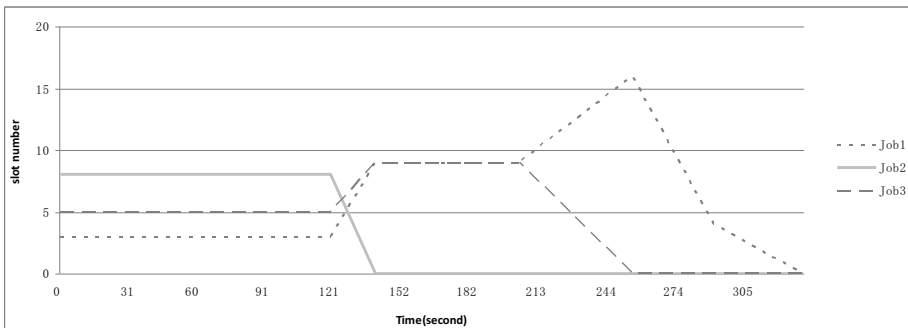


Fig. 1. The number of slot allocated to the three job changes over time

Table 2 shows the running time, cost and utility of the three jobs.

Table 2. Information of jobs

Job name	cost	running time(second)	utility
Job1	83.6	310	183.76
Job2	118	140	7.2
Job3	96	254	40.35

7 Conclusion and Future Work

In this paper, we presented a scheduling algorithm in hadoop environment. Assume that the user focus on satisfaction, rather than just time or cost. The proportion of the allocation of resources is determined by the user bid which determined by the system. The user's preferences will affect their charges. Algorithm charge more to reduce the job running time; to the contrary, extend the running time to reduce costs to the users. Price also will float with the utilization of the resources which making the system a high satisfaction.

In future work, we will consider the satisfaction both system and individual job and solve it using game theory.

References

1. Armbrust, M., et al.: Above the Clouds: A Berkeley view of cloud computing. University of California at Berkeley (2009)
2. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. *Commun. ACM* 51(1), 107–113 (2008)
3. Mingchu, L., et al.: Grid Resource Allocation Model Based on Incomplete Information Game. *Journal of Software* 23(2), 428–438 (2012)
4. Cong, W., et al.: Resource Allocation Model to Achieve Optimize Profit in Virtual Data Center. *Journal of Northeastern University* 32(11), 1546–1549 (2011)
5. Hadoop CapacityScheduler, http://hadoop.apache.org/common/docs/current/capacity_scheduler.html
6. Hadoop FairScheduler, http://hadoop.apache.org/common/docs/current/fair_scheduler.html
7. Sandholm, T., Lai, K.: Dynamic Proportional Share Scheduling in Hadoop. In: Frachtenberg, E., Schwiegelshohn, U. (eds.) *JSSPP 2010. LNCS*, vol. 6253, pp. 110–131. Springer, Heidelberg (2010)
8. Xicheng, D., Ying, W., Huaming, L.: Scheduling Mixed Real-time and Non-real-time Applications in MapReduce Environment. In: *17th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, Tainan, pp. 9–16 (2011)
9. Polo, J., et al.: Performance-Driven Task Co-Scheduling for MapReduce Environments. In: *2010 IEEE/IFIP Network Operations and Management Symposium - NOMS*, pp. 373–380 (2010)
10. Kc, K., Anyanwu, K.: Scheduling Hadoop Jobs to Meet Deadlines. In: *Proceedings of the 2010 IEEE 2nd International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 388–392 (2010)
11. You, H., Yang, C., Huang, J.: A load-aware scheduler for MapReduce framework in heterogeneous cloud environments. In: *Proceedings of the ACM Symposium on Applied Computing*, pp. 127–132 (2011)

Research and Application of Regression Test Cases Aided Design Method Based on Association-Mode

Wenhong Liu^{*}, Xin Wu, Yuheng Hao, and Jimao Wang

Beijing Institute of Tracking and Telecommunications Technology, Beijing, China
lwhsteel@163.com

Abstract. Inadequate regression test case design after software changes often leads to poor effectiveness. In order to solve this problem, this paper proposes a novel design method based on the association-mode. Through analyzing the relationship between the software testing requirements, test items, test cases, software failures and software changes, the Influence domain analysis of software changes is completed by using these association-mode. This paper implements the filtering and the aiding design of the regression test cases, and developed a test case aiding design tool by considering the relationship of the testing requirements and the priority of the test cases. Test cases from practical projects show that this method can effectively detect the failures. This regression test cases aided design method can help the testers to design regression test cases effectively and avoid the problems of poor effectiveness as a result of the inadequate regression test case design.

Keywords: regression testing, testing requirements, test case, association-mode.

1 Introduction

Software testing is an important means to improve software quality. Developing software dynamic testing and regression testing of the changed software is a very important means especially for the large-scale software systems which include a variety of operating modes. For the iterative software development model which is widely used, continuous regression testing is needed much more [1]. There are many examples of fatal failure caused by inadequate regression testing. For example, the reason of the failure of the Ariane 5 rocket launch is the ineffective regression testing for the reuse code. Therefore, it is necessary to improve the adequacy and effectiveness of the software regression testing.

Software regression test is used to re-test the software system which is modified or extended (such as system enhancements or upgrades), and ensure the modification while not introduce new bugs. Software changes are inevitable in the process of the software development. And they may be caused by the following reasons:

1. Corrective changes. To correct the errors found in the verification confirmation process such as test. Including the changes caused by the correction of design errors, program errors, data errors and documentation errors.

^{*} Corresponding author.

2. Adaptive changes. To adapt the change of the software running environment, such as hardware configuration changes, data format or file structure changes, software support environment changes.
3. Improvable changes. The changes for expanding functionality or improving performance, such as expanding functionality, improving performance and making the maintenance easier.
4. Preventive changes. The changes to prevent problems from occurring.

The following problems may exist in the changes for the software fault:

1. Change itself is not correct.
5. The incompleteness of the change caused by the ineffective influence domain analysis.

The following problems may exist in other categories of changes:

1. There is a failure in the new code itself.
6. The influence on the original codes.

Therefore, when performing regression testing, it not only needs to test the change part to prove that the modification is correct, but also need to add new test cases to verify the new code. In addition, it needs to select some of the existing test cases to prove that no new bugs are introduced in the unchanged part.

Generally, there are two software regression testing strategy: the first is to re-run all test cases again; and the second is to analyze the influence domain based on the software changes, then design the regression test cases. The advantage of the first strategy is that more influenced parts can be found in the complete test, the disadvantage is that it takes much time and large resources in this method, and generally, no software development organization can afford it. The second strategy has the advantage of less cost, and the disadvantage is the poor accuracy of qualitative analysis, leading to the lack of guarantee of the validity and the adequacy of the test. In summary, we use the second strategy in the regression testing generally. However, aided support means for software regression test cases design need to be provided to improve its effectiveness and adequacy.

Currently, there are a lot of studies on the effectiveness and the adequacy of software testing. Test case suite generation and reduction technology in the regression testing has become one of the hot research fields of software engineering [2-7]. However, these methods have some limitations in the test applications based on the software requirements specification.

An aiding method for regression test cases design in accordance with the requirements specification was proposed in this paper. The regression testing strategies work based on the influence domain analysis of software changes. By analyzing the association-mode between the software testing requirement, test items, test cases, software failures and software changes, we found the method to identify the software regression test cases. And after considering the association-mode, we got the aiding design method for the regression test cases.

2 The Association-Mode in the Software Testing Process

2.1 The Association-Mode between the Test Requirement

The set of the software testing requirements is defined as $R = \{ r_1, r_2, r_3, \dots, r_i, \dots \}$, which is needed to be covered by the tested software systems. The testing requirement can be a statement or a function (process or method) from the perspective of structure, and it can be a function or measured characteristics from the perspective of function [1].

During the testing requirements analysis, the testing requirements are determined according to the software requirements specification. The testing requirements include software functionality, performance, interface, GUI, security, reliability and testability.

There exist relationships between the testing requirements [8]. During the testing requirements analysis, the association-mode between the software requirements can be obtained according to the software requirements specification. In the software requirements specification described by the structured method, the relationship between the software requirements can be determined according to the software data flow diagram and control flow diagram. In the software requirements specification described by the object-oriented method, the association-mode can be determined according to the use case diagram [9-10]. Considering the convenience in practice, the association-mode between the software testing requirements can be expressed as an undirected graph.

2.2 The Association-Mode between the Test Requirement and Test Item

The set of the software test item is defined as $I = \{ item_1, item_2, item_3, \dots, item_j, \dots \}$, which is determined by the tested software systems. The test items are designed to implement a test requirement effectively. The description of the test items includes: name, sign, details of test items, test methods, test adequacy requirements, termination requirements of the test items, priority, and the association between the test items and the testing requirements.

By analyzing the association-mode between the test items and the testing requirements, a testing requirement is covered by one or more test items.

2.3 The Association-Mode between the Test Item and Test Case

The set of the software test cases is defined as $T = \{ t_1, t_2, t_3, \dots, t_k, \dots \}$, which is designed for testing software. By analyzing the association-mode between the set of test cases and the set of test items, it is concluded that a test case may cover one or more test items, and a test item may require one or more test cases to complete testing requirements.

2.4 The Association-Mode between the Test Case and Software Failure

The set of the software failures is defined as $F = \{ f_1, f_2, f_3, \dots, f_i, \dots \}$, which is identified after executing the software test, and the failures are the external embodiment of the software defect. Thus, a software failure must be associated with a test case. In other words, a software failure found during the test must be generated by executing a test case. In this way, we can find the corresponding test case by a failure to a certainty.

2.5 The Association-Mode between the Software Failure and Software Change

The set of the software change is defined as $C = \{ c_1, c_2, c_3, \dots, c_m, \dots \}$, including the modifications for software failures, adaptive changes, improvable changes and preventive changes. For software changes, not only re-executing the un-passed test case is needed, but also selecting the existing test cases or adding new test cases according to the results of the influence domain analysis. Generally, new test cases need to be added to the new testing requirements.

After analyzing the association-mode between software failures and software changes, we can get the following conclusions: a software failure is associated with one or more software changes, and a software change can be associated with one or more software failures. Considering the demand for the changes caused by new software requirement or adaptive change requirement, and the convenience of implement as well, we can describe the software change is associated with zero software failure in this condition.

The association-mode during the software testing process is shown in Fig. 1.

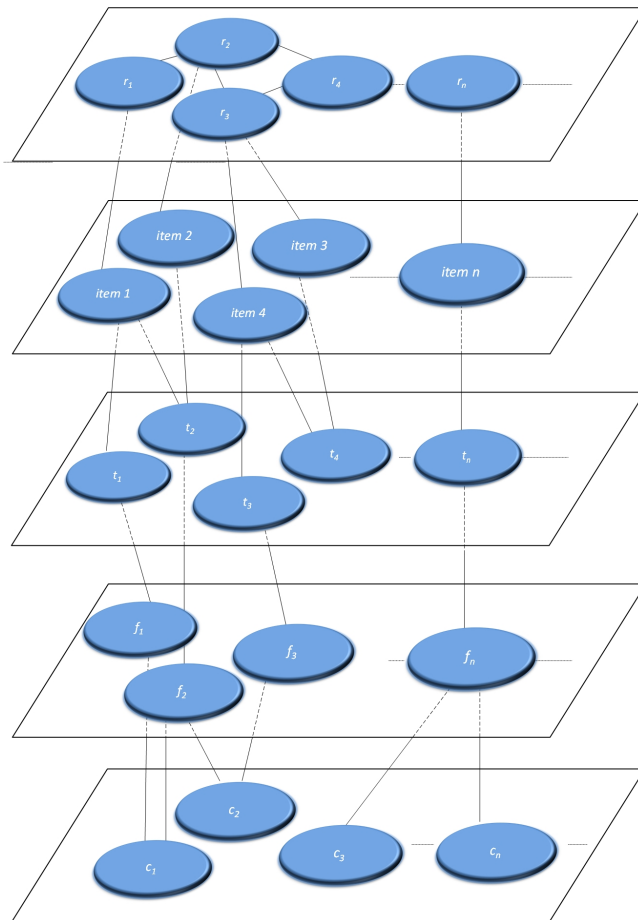


Fig. 1. Tracing relationship diagram in the software testing

3 Regression Test Cases Design Method Based on Association-Mode

3.1 Regression Testing Process Analyzing

By analyzing the software regression test, we can get its process as shown in Fig. 2.

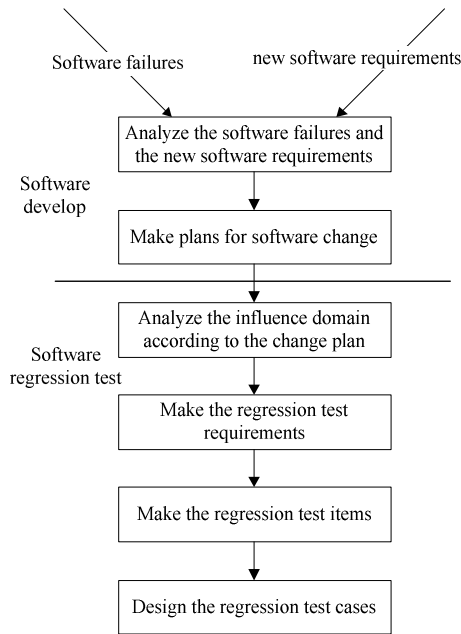


Fig. 2. Software regression testing process

Through analyzing the new software requirements and the software failures found during the test, software developers work out the software change plan. And software testers analyze the influence domain based on it, then make the regression testing requirements clear, get test items and design test cases.

Testers tend to verify the corrections of the modifications in the regression testing process, while ignoring whether the changes make some functions invalid. In the same way, only the modified itself is considered in the design of regression test cases, while the other influenced part is ignored.

Therefore, in the study of regression test cases design methods, we need to consider this part much more.

3.2 Related Definitions

By analyzing the association-mode in the software testing process, we can get the following definitions:

Definition 1: the original testing requirements set is defined as R , test items set as I , test cases set as T ; in regression test, testing requirements set is defined as the R' , test items set as I' , test case set as the T' ; test failures set is defined as F , test changes set as C .

Definition 2: the attributes of the software testing requirements r_i include: name, sign, description.

Definition 3: the attributes of the test item $item_j$ include: name, sign, description, test methods, test adequacy requirements, termination requirements of the test items, priority, and the association between the test $item_s$ and the testing requirements. Among them, the priority set is: high, middle, low.

Definition 4: the attributes of software test case t_k include: name, sign, description, priority, association with the test items, initialized requirements, premise and constraints, test procedure, the expected results and evaluation criterion. Among them, the test priority set is: high, middle, low.

Definition 5: the attributes of software failure f_i include: sign, failure categories, failure level, associated test cases, failure description and suggestions for improvement.

Definition 6: the attributes of software changes c_m include: sign, description, change categories which include corrective changes, adaptive changes, improvable changes and preventive changes.

3.3 The Algorithm

According to the influence domain analysis, the aiding design algorithm of the software regression test cases is described as follows:

1. When the change is corrective change, the following steps should be executed for the aiding design of the software regression test cases;
 - (a) Search the attribute of software changes to find the software failures associated with it, and put the associated test case into regression test case set T' ;
 - (b) Traverse the attributes of all test cases in T' , and find the associated testing items according to the relationship between the test cases and test items, then put them into I' ;
 - (c) Traverse all the test items in I' , and find the associated testing requirements according to the relationship between the test items and testing requirements, then put them into R'' ;
 - (d) Traverse all the requirements in R'' , put all testing requirements associated with the requirements in R'' into testing requirements set R' , and put all requirements in R'' into R' too.
 - (e) Traverse all the requirements in R' , put all test items associated with R' and have 'High' priority into I' .
 - (f) Traverse all the test items in I' , put all test cases associated with I' and have 'High' priority into T' .
- 2 When the change is not corrective change, put the requirements of the software change into the software requirements set R' .

The filtering result of the existing test cases in regression testing can be got by following the steps (a) to (f) in case 1.

The test case design wizard can import the new testing requirements in case 2, and help the testers design the new test items and the new test cases. It realizes the function of wizard in regression test case designing.

3.4 Algorithm Implementation

The class diagram of the test case design aiding tool is shown as Fig. 3. It consists of four entity classes and four associated classes which connect the entity classes.

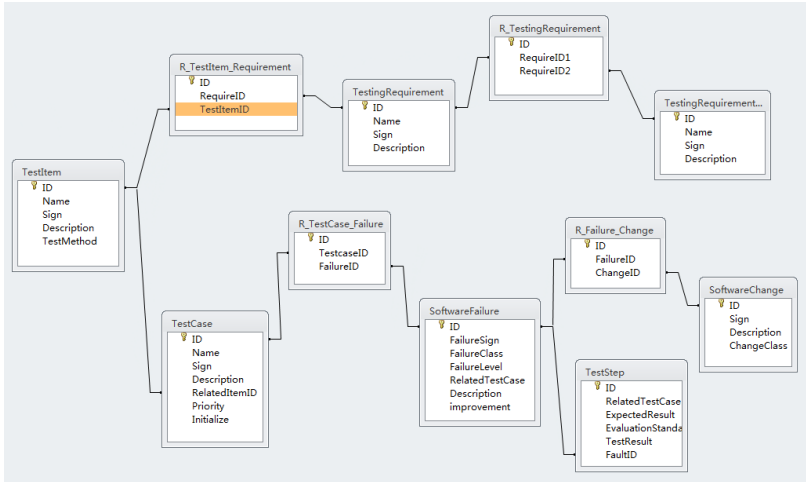


Fig. 3. The class diagram of the test case design aiding tool

The software screenshot is shown as Fig. 4.

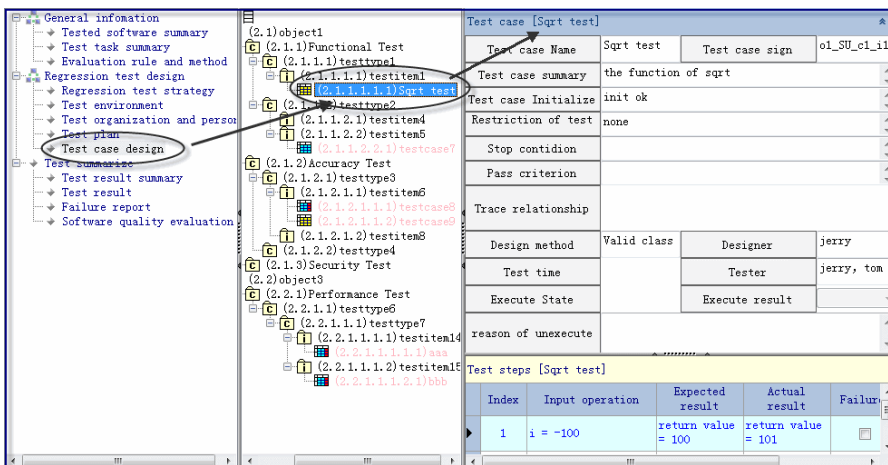


Fig. 4. The software screenshot of the tool

4 Application Examples

Based on the analysis of data transmission software requirements, get the relationship as shown in Fig. 5. A total of 168 test cases have been designed for the 8 functional modules.

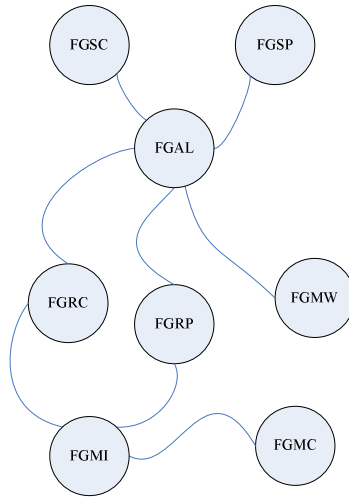


Fig. 5. The association between testing requirements

Five software failures have been found in the software testing process. In regression testing, the testers only selected the 6 test cases which are not passed for regression testing, and found no new software failures. But the failure that the interface does not match was found while the software was running. After analyzing, it found that there were some influences to the processing functions of the FGAL when modify the FGRC data type. And software testers failed to take this problem into account, which resulted in the insufficiency in the regression testing.

By using the regression test case generation algorithm introduced in this paper, the relevant test cases which related to FGRC were also added to the regression test cases set, and got 11 more regression test cases. The coverage of test cases to testing requirements is shown in Table 1. After performing the test, it can find the interface unmatched failure caused by the modifications of FGRC. It improves the adequacy of the software regression testing to a certain degree.

Table 1. The tracking association between the Regression testing requirements and test cases

Regression testing requirements	Regression test cases
FGMI	1
FGMC	2
FGRC	7
FGAL	1

In the past year, 156 software testing projects have been tested with this method in Software Testing Center, and the failure caused by inadequate regression testing did not happen again.

5 Conclusion

Through the analysis of the association-mode between the test changes and test failures, test failures and test cases, test cases and test items, test items and testing requirements, and between testing requirements itself, a regression test cases aiding design method was proposed in this paper, and a test case aiding design tool was developed. This method has been applied in the regression testing of the data transmission software testing project. And the results show that it can improve the adequacy, validity and the degree of automation in regression test cases design effectively.

References

1. Gu, Q., Tan, G.B., Chen, D.X.: A Test Suite Reduction Technique for Partial Coverage of Test Requirements. *Chinese Journal of Computers* 34(5), 779–887 (2011)
2. Harrold, M.J., Gupta, R., Soffa, M.L.: A Methodology for Controlling the Size of a Test Suite. *ACM Transactions on Software Engineering and Methodology* 2(3), 270–285 (1993)
3. Chen, T.Y., Lau, M.F.: A New Heuristic for Test Suite Reduction. *Information and Software Technology* 40(5), 347–354 (1998)
4. Jeffrey, D., Gupta, N.: Improving Fault Detection Capability by Selectively Retaining Test Cases During Test Suite Reduction. *IEEE Transactions on Software Engineering* 33(2), 108–123 (2007)
5. Black, J., Melachrinoudis, E., Kaeli, D.: Bi-criteria Models for All-uses Test Suite Reduction. In: *Proc. of International Conference on Software Engineering*, Edinburgh, Scotland, UK, pp. 106–115 (2004)
6. Cai, S.M., Mei, D.H.: Regression Test Case Generation Based on Dynamic Slicing And UML Diagram. *Computer Engineering* 35(8), 70–72 (2009)
7. McMaster, S., Memon, A.M.: Call-stack Coverage for GUI Test Suite Reduction. *IEEE Transactions on Software Engineering* 34(1), 992115 (2008)
8. Chen, Y.M.H., Wang, L., Zhu, G.H.: Regression Test Case Reduction Model Based on Association Mode. *Computer Engineering* 37(2), 63–68 (2011)
9. Wang, X.: Application of UML Statechart Diagram in Regression Test. *Computer Engineering* 35(4), 63–65 (2009)
10. Chen, S.F., Zheng, H.Y.: Dependence Analysis And Regression Testing of Object Oriented Software. *Journal of Computer Applications* 39(11), 3110–3113 (2009)

Virtual Network Embedding Algorithm Based Connective Degree and Comprehensive Capacity

Ying Yuan¹, Cuirong Wang², Na Zhu³, Cong Wan², and Cong Wang²

¹ School of Information Science and Engineering,
Northeastern University, Shenyang, 110004, China
yuanying1121@gmail.com

² School of Northeastern University at Qinhuangdao, 066004, China
wangcr@mail.neuq.edu.cn, Congw1981@gmail.com

³ School of Qinhuangdao Institute of Technology, 066004, China
zhuna81@126.com

Abstract. Network virtualization allows multiple virtual networks to coexist on a shared physical substrate infrastructure. As far as possible to support more virtual networks on a shared substrate network, efficient physical resource utilization is crucial. This paper presents a novel approach to increase utility of the substrate network. As an optimization problem, such virtual network embedding problem is formulated as an integer linear programming model. By introducing a sliding window with priority for VN requests, the algorithm embeds virtual nodes based on both connective degree and comprehensive capacity. Experimental results show that comparing with the existing VNE algorithm the proposed algorithm achieve higher VNs accept ratio and gain higher revenue-cost ratio for substrate network.

Keywords: network virtualization, embedding algorithm, virtual network, connective degree, sliding window.

1 Introduction

Network virtualization is an important method to solve ossification problem of the current Internet [1]. Network virtualization allows multiple heterogeneous virtual networks (VNs) to coexist on a shared infrastructure [2]. In the network virtual environment, internet service provider (ISP) is divided into two parts: infrastructure provider (InP) and service providers (SP). InPs manage the physical infrastructure while SPs create VNs and offer end-to-end services [3].

One of the major challenges in network virtualization is the efficient allocation of substrate resources to the Virtual Networks (VNs)--a problem known as Virtual Network Embedding (VNE) [4]. Virtual network Embedding problem is known to be NP-hard even in the offline case. Even if all the virtual nodes are mapped, it is still NP-hard to mapped virtual links [5]. In order to solve the VN embedding problem, a number of heuristic algorithms have appeared in the relevant literature [2-9]. Without reducing any problem space, heuristic algorithms can be divided to two categories:

one-stage VNE algorithm and two-stage VNE algorithm. The VN embedding problem can be divided into two phases: node mapping phase and link mapping phase. The former considers the constraints of both virtual nodes and links at the same time in the VNE process. The latter preselects all the substrate nodes satisfying the requirements of virtual nodes in a greedy way, and after that, searches the substrate network to find loop-free paths meeting the needs of virtual links. In the node mapping phase, these early studies have primarily focused on considering the resource for these virtual nodes without the degree of node. Therefore, we study the problem how to optimally configure the VN request base on the degree of node.

Backtracking algorithms in conjunction with graph isomorphism detection to map VN requests are used in a dynamic fashion in [6]. The algorithm maps nodes and links during the same stage. The advantage of this single stage approach is that link mapping constraints are taken into account at each step of the mapping. When a bad mapping decision is detected it can be revised by simply backtracking to the last valid mapping decision, whereas the two stage approach has to remap all links which is very expensive in terms of runtime. The authors in [7] advocate a different approach: rethinking the design of the substrate network to enable simpler embedding algorithms and more efficient use of resources, without restricting the problem space. The authors in [8] propose VNE algorithm D-ViNE and R-ViNE. They formulated the VN embedding problem as a mixed integer program through substrate network augmentation, and then relaxed the integer constraints to obtain a linear program. The authors in [5] propose a new greedy virtual network reconfiguration algorithm VNR to minimize the number of overloaded substrate links as well as reducing the cost of reconfiguration. The authors in [9] use a joint node and link mapping approach for the VI mapping problem and develop a virtual infrastructure mapping algorithm.

In this paper, we will propose a new virtual network embedding algorithm, denoted by WD-VNE. We note that the connective degree of nodes could have significant impact on the success and efficiency of embedding VNs. We adopt the sliding window technique for VN request with look-ahead capabilities. The VN requests within the sliding window are sorted based on their revenues, so that the VN with the maximum revenue is considered first for embedding. We sort the virtual nodes by degree in descending order and by comprehensive capacity in descending order when degree is equal. Then we map the highest degree node firstly to physical node.

Rest of the paper is organized as follows. In Section, we give the detailed description of VNE model. In section 3, we discuss how to deal with VNE base on connective degree and comprehensive capacity. The simulation environment and results are given in Section 4. Section 5 gives the conclusion.

2 Network Model and Problem Description

In this section, we first provide the network model, including substrate network, virtual network request. The proposed solution represents virtual as well as substrate networks as undirected graphs. We then introduce the virtual network embedding problem, including basic definitions and formulations.

2.1 Substrate Network

The substrate network will be modeled $P = (\hat{h}^p, \zeta_h^p, \ell^p, \zeta_\ell^p)$, where \hat{h}^p and ℓ^p denote the set of substrate nodes and substrate link, respectively. We denote the set of loop-free substrate paths by \wp^p . The notations ζ_h^p denote the attributes of the substrate nodes. The notations ζ_ℓ^p denote the attributes of the substrate path. In this paper, each substrate node $h^p \in \hat{h}^p$ is associated with the CPU capacity. Each substrate link $l^p(m, n) \in \ell^p$ between two substrate nodes m and n is associated with the bandwidth.

2.2 Virtual Network Request

Similar to the substrate network, the virtual network will be modeled $V = (\hat{h}^v, \zeta_h^v, \ell^v, \zeta_\ell^v)$, where \hat{h}^v is the set of substrate nodes and ℓ^v is the set of substrate link. The notations ζ_h^v and ζ_ℓ^v denote the constraints of virtual nodes and edges, respectively. We denote a virtual network request can be defined by $VNR = (V, T_a, t)$, where T_a is the arrival time, t the duration of VN staying in the substrate network.

2.3 Virtual Network Embedding

The virtual network embedding problem can be defined as an embedding action ∇ from a virtual network request VNR to a subset of the substrate network P. The subset is the image of V in P, so it should satisfy the constraints $(\zeta_h^p, \zeta_\ell^p)$ from P. The embedding action can be expressed as follows:

$$\nabla : V \mapsto (\hat{h}^{p^*}, \psi_h, \ell^{p^*}, \psi_\ell) \quad (1)$$

where $\hat{h}^{p^*} \subseteq \hat{h}^p$, $\ell^{p^*} \subseteq \ell^p$.

It can be further detailed as node embedding and link embedding:

$$\nabla_h : (\hat{h}^v, \zeta_h^v) \mapsto (\hat{h}^{p^*}, \zeta_h^p) \quad (2)$$

$$\nabla_h(h^v) \in \hat{h}^{p^*}, \forall h^v \in \hat{h}^v \quad (3)$$

$$Rcpu(h^v) \leq Ccpu(\nabla_h(h^v)) \quad (4)$$

$$dis(loc(h^v), loc(\nabla_h(h^v))) \leq \tau \quad (5)$$

$$\nabla_\ell : (\ell^v, \zeta_\ell^v) \mapsto (\ell^{p^*}, \zeta_\ell^p) \quad (6)$$

$$\nabla_\ell(l^v) \in \ell^{p^*}, \forall l^v \in \ell^v \quad (7)$$

$$Rbw(l^v) \leq Cbw(\nabla_\ell(l^v)) \quad (8)$$

$$Cbw(\nabla_\ell(l^v)) = \min_{\wp^p \in \wp^{p^*}} Cbw(l^p)$$

where $Rcpu(h^v)$ is requested CPU constraints for the virtual node h^v , $Ccpu(\nabla_h(h^v))$ is the available CPU capacity of substrate node, $Rbw(l^v)$ is requested bandwidth constraints for the virtual edge l^v , $Cbw(\nabla_\ell(l^v))$ is the available bandwidth capacity of substrate path. The available bandwidth capacity of a substrate path $\wp \in \wp^{p^*}$ is the minimum bandwidth capacity of substrate edge in this path.

Fig. 1 shows an example of VNE problem. Figure 1(a) and Figure 1(b) shows two VN request. Where the numbers in rectangles next to the virtual nodes represent the amount of node resources requested by the nodes and the numbers next to the virtual edges represent the edge resource constraints of the edges. There are many different embedding solutions for the same problem. Figure 1(c) and Figure 1(d) shows two different embedding solutions.

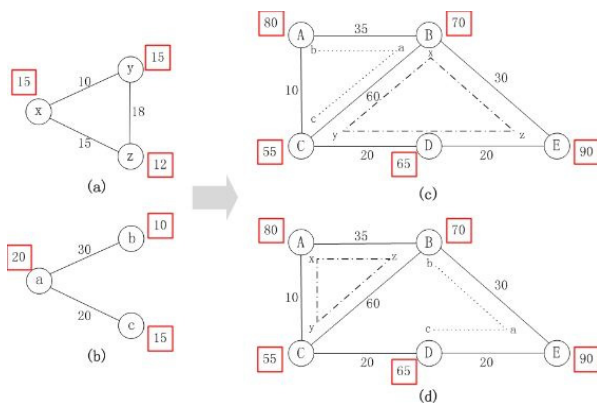


Fig. 1. Embedding of VN requests onto a shared substrate network

2.4 System Object

The main objective of virtual network embedding is to maps the virtual network requests to the substrate network to make efficient use of the substrate network resources.

Similar to the early work in [2, 7, 8, 9], revenue is defined the sum of total an virtual network request gain from InP. It is one of the main factors we use to judge the performance of an embedding algorithm. It is defined as

$$R(V) = \alpha_R \cdot \sum_{l^v \in l^v} bw(l^v) + \beta_R \cdot \sum_{h^v \in h^v} cpu(h^v) \quad (10)$$

where $bw(l^v)$ is the bandwidth of the request link, $cpu(h^v)$ is the capacity of the request node. α_R and β_R are weighting coefficients to balance the effect of bandwidth and CPU.

The cost of embedding a virtual network request is defined the sum of total substrate resources allocated to that virtual network following as

$$C(V) = \alpha_C \cdot \sum_{l^v \in l^v} bw(l^v) Length(\varphi) + \beta_C \cdot \sum_{h^v \in h^v} cpu(h^v) \quad (11)$$

where $Length(\varphi)$ is the hop count of the virtual link l^v when it is assigned to a set of substrate links. α_C and β_C are weighting coefficients similarly to α_R and β_R .

We define the virtual network request acceptance ratio as:

$$AcceptRatio = \lim_{T \rightarrow \infty} \frac{\sum_{t=0}^T VNRA}{\sum_{t=0}^T VNR} \tag{12}$$

where $VNRA$ is the number of virtual network request successfully accepted by the substrate network, and VNR is the number of the arrival virtual network request.

The long-term revenue to cost ratio indicates the efficiency of substrate network resource usage, which is an important factor to judge the performance of a virtual network embedding algorithm. We define the long-term revenue to cost ratio as

$$R/C = \lim_{T \rightarrow \infty} \frac{\sum_{t=0}^T R(V(t))}{\sum_{t=0}^T C(V(t))} \tag{13}$$

3 WD-VNE Algorithms

3.1 System Model

There are several different ways to embed a given virtual network request into the substrate network. How to guarantee optimal resource usage that is the core problem of virtual network embedding. We model the problem of embedding of a virtual network as a mathematical optimization problem using linear programming (LP). The objective of our optimization problem is defined as follows:

Objective Function

$$Minimize \sum_{\varepsilon \in \ell^v} \sum_{l \in \wp^p} |l| bw(l, \varepsilon) \tag{14}$$

Constraint

$$\forall i, j \in h^p, \wp^{ij} \in \wp^p, \forall w \in \ell^v \quad Cbw(\wp^{ij}) = \min_{l^p \in \wp^p} Cbw(l^p) \geq Rbw(w) \tag{15}$$

$$Cbw(l^p) = Bw(l^p) - \sum_{l^v \rightarrow l^p} Bw(l^p) \tag{16}$$

where $|l|$ is the length of substrate path; $bw(l, \varepsilon)$ is the amount of bandwidth allocated on path l for virtual link ε ; $Cbw(\wp^{ij})$ is the available bandwidth capacity of a substrate path from i to j ; $\sum_{l^v \rightarrow l^p} Bw(l^p)$ is the total amount of bandwidth capacity allocated to different virtual links hosted on the substrate link; $Bw(l^p)$ is the total amount of bandwidth capacity of the substrate link.

3.2 WD-VNE Algorithm Description

In this section, firstly, we consider the connective degree of node. The comprehensive capacity for node is defined by

$$\mathbb{Z}(h) = cpu(h) \cdot \sum_{l \in \ell(h)} bw(l) \oplus \sqrt{D \cdot \sum_{h^{nb} \in h} (cpu(h^{nb}) \cdot \sum_{l \in \ell(h^{nb})} bw(l))^2} \tag{17}$$

where D is the degree of node h ; h^{nb} is the immediate neighboring nodes of h .

For VN request, we adopt the sliding window technique. We define a window w as a set of VN requests. Initially, we use window w to select VN requests from the beginning of the VN request queue. The window has a maximum waiting period T_w . For each VN request, we set the maximum mapping count. The details of the sliding window algorithm are shown as follows:

Algorithm 1. The algorithm for sliding window $SW(VNRQ)$

1. Initialize w = Size of Sliding Window, m =Maxima mapping Count;
2. Select w VN requests from the beginning of the $VNRQ$;
3. $EQList \leftarrow$ sorted w VN requests according to their revenues;
4. Process VN requests one by one in $EQList$ using VN embedding algorithm;
5. For sliding window, if its maximum waiting period is expired, sliding window by $W_{inc} = w - |w|$ (where $|w|$ is the number of VN requests have not been mapped in previous window, i.e. in step 4.);
6. For each VN request in the window, if its maximum embedding threshold which indicates the maximum mapping times for one VN request is reached, the VN request will be put into a waiting queue;
7. $VNRQ \leftarrow$ the waiting queue; Go to step 2.

WD-VNE algorithm takes equation (14) as objective function. It is a backtracking algorithm. In each round, the algorithm will sort the virtual nodes by degree in descending order and by comprehensive capacity in descending order when degree is equal. Then maps the highest degree node firstly to such physical node, which degree is equal to the virtual node and its comprehensive capacity is highest in same degree physical nodes of the substrate network. The details of the WD-VNE algorithm are shown as follows:

Algorithm 2. The WD-VNE algorithm $WDVNE()$

1. $Q \leftarrow H$, where H is the virtual node which has highest degree in VNR , Q is a FIFO queue;
2. $Embedding(H, n_H)$, where n_H is a counter to calculate backtracking times for virtual node H , for all virtual node if there is the first time to map, $n_i = 0$;
3. $Q_{Temp} \leftarrow$ Sort all H 's unmapped immediate neighbor according to degree of descending order, when degree is equal, order by $\mathbb{Z}(h)$ (equation 17) in descending order;
4. For each node K in Q_{Temp} , $check_node_link_mapping()$, if has a solution, add all K 's immediate neighbors to Q ; otherwise $n_H = n_H + 1$, go to step 2.

Algorithm 3. $Embedding(H, n_H)$

(H is a virtual node, n_H is backtracking times counter)

1. $Total_backtracking_times + 1$. If $Total_backtracking_times > Max$ backtracking times, break and let algorithm 2 returns non solution for this VN;
2. Calculate H 's candidate physical nodes set P_H ;
3. Sort P_H according to degree of ascending order, when degree is equal, order by $\mathbb{Z}(h)$ in descending order;

4. Select the n_H th candidate physical nodes. If it exists, return the solution else backtrack to H 's upper level neighbor and re-embed the neighbor in Algorithm 1, i.e. do a cross level backtracking;
5. If H is the highest degree node of the VN request and there is no n_H th candidate physical nodes, let algorithm 2 returns non solution for this VN and process next VN request.

In the algorithm 2 we use function *check_node_link_mapping*() to pick one substrate node from the candidate neighbor list of H 's physical position and check if the mapping is feasible according to CPU, link bandwidth and physical link hops constraints.

4 Simulation

We implemented the WD-VNE algorithm using the CloudSim3.0.1 on a PC which has one Intel Core i7-3770 CPU and 20G DDR3 1600 RAM. We write a random topological generator in java to generate topologies for the underlying substrate networks and virtual networks in CloudSim. We simulated 1000 VN request for each algorithm, each test run 20 times. All the results presented for an experiment are an average of 20 runs of simulation.

Table 1. Parameters in simulation

Topology:	Substrate Network	Virtual Network
Number of Nodes:	80	4-8
Connectivity:	0.2	0.5
Node Capacity:	100 unit	3-30 unit uniform distribution
Bandwidth Capacity	100 unit	3-30 unit uniform distribution

As shown in Fig.2, the acceptance ratio of the WD-VNE algorithm was significantly larger than that of the vnmFlib algorithm. The reason is that the algorithm embed virtual node to physical node on the basis of connective degree and comprehensive capacity, such hybrid mechanism can make substrate network to accept more virtual network requests at the same time than other approaches.

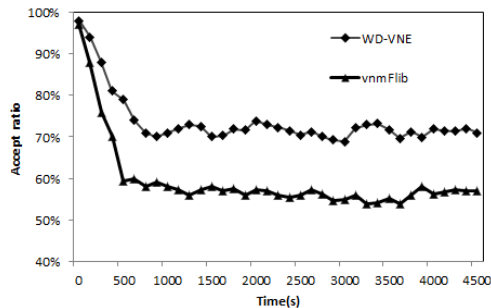


Fig. 2. Accept ratio

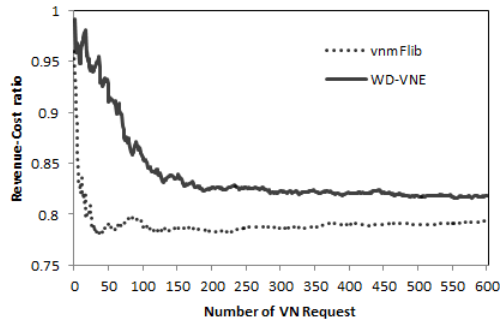


Fig. 3. Revenue-Cost ratio

From Fig. 3 we can see that the Revenue-Cost(R/C) ratio of the WD-VNE algorithm was larger. Along with the increase of the number of virtual network, their R/C ratio becomes more balanced. The reason is that the WD-VNE algorithm adopted the sliding window technique and considers the connective degree of node. The higher R/C ratio indicates higher resource utilization, which further results in accepting more VNs at certain infrastructure resources.

5 Conclusions

In network virtualization, the problem of efficiently allocating network resources of multiple VNs coexist on a shared substrate network is challenging. This paper introduced a novel VNE algorithm which introduce sliding window technique for VN request queue and embed the virtual nodes based on both connective degree and comprehensive capacity. Simulation results show that our algorithms outperform the previous approaches in terms of the accept ratio and R/C ratio.

Acknowledgments. This work was supported by The Central University Fundamental Research Foundation, under Grant. N110323009.

References

1. Larry, P., Scott, S., Jonathan, T.: Overcoming the Internet Impasse Through Virtualization. *J. Computer* 38(4), 34–41 (2005)
2. Mosharaf, C., Muntasir, R.R., Raouf, B.: Vineyard: Virtual Network Embedding Algorithms with Coordinated Node And Link Mapping. *J. IEEE/ACM Transactions on Networking* 20(1), 206–219 (2011)
3. Rahman, M.R., Aib, I., Boutaba, R.: Survivable Virtual Network Embedding. In: Crovella, M., Feeney, L.M., Rubenstein, D., Raghavan, S.V. (eds.) *NETWORKING 2010*. LNCS, vol. 6091, pp. 40–52. Springer, Heidelberg (2010)
4. Sarang, B.M., Raghavan, S.V.: Simulated Annealing Algorithm for Virtual Network Reconfiguration. In: *2012 8th EURO-NGI*, pp. 95–102. IEEE Press, Karlskrona (2012)

5. Ilhem, F., Nadjib, A., Guy, P., Hubert, Z.: VNR Algorithm: A Greedy Approach for Virtual Networks Reconfigurations. In: GLOBECOM 2011, pp. 1–6. IEEE Press, Houston (2011)
6. Jens, L., Holger, K.: A Virtual Network Mapping Algorithm Based on Subgraph Isomorphism Detection. In: The 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures (VISA 2009), pp. 81–88. ACM, New York (2009)
7. Minlan, Y., Yung, Y., Jennifer, R., Mung, C.: Rethinking Virtual Network Embedding: Substrate Support for Path Splitting and Migration. *J. ACM SIGCOMM CCR* 38(2), 17–29 (2008)
8. Mosharaf, C., Muntasir, R.R., Raouf, B.: Virtual Network Embedding with Coordinated Node And Link Embedding. In: INFOCOM 2009, pp. 783–791. IEEE Press, Rio De Janeiro (2009)
9. Hong, F.Y., Vishal, A., Chun, M.Q., Hao, D., Xue, T.W.: A Cost Efficient Design of Virtual Infrastructures With Joint Node and Link Mapping. *Journal of Network and Systems Management* 2012 20(1), 97–115 (2012); Deep, M. (ed.)

Fire Prevention in Gas Stove with Gas Sensors

Soojung Kang, Bitna Song, and Ilhoon Shin *

NowonGu GongleungDong, Seoul National University of Science and Technology,
3th Building 207, Seoul, South Korea, 139-743
ilhoon.shin@snut.ac.kr

Abstract. This paper presents to use gas sensors to implement a safer home from a fire caused by the gas stove. Nowadays, the ratio of the families that consist of one person or old people only is increasing and thus the probability of the fire caused by forgetting turning off the gas stove is also increasing. In order to prevent this kind of fire, the presented appliance detects the poisonous gases generated on burning of pots or food, and it closes the gas valve automatically using a motor on the detection of the emergent situation. We implement the presented idea to the prototype appliance and prove the correctness of its function.

Keywords: Gas Sensors, Fire Prevention, Smart Home.

1 Introduction

Recently, as various kinds of sensing devices have appeared and their prices are getting cheaper, building a smart home that automatically controls home appliances using sensors and processors has been presented [1-3]. For example, home appliances are wirelessly connected and they can be controlled by smart phones [1]. This work also presents to build a safer home from a fire using gas sensors.

Every home uses a gas stove. Gas stoves can be a source of a fire and they should be carefully managed. There are two main cases of the fire caused by the gas stove. The first case is a gas leak. If there is the gas leak because of problems of the gas stove or of the gas hose, a small ember can lead to a serious fire. Therefore, there was previous research that detects the gas leak and blocks the gas valve automatically [4]. It can prevent the fire from the gas leak, and however it does not detect the second case of the fire caused by the gas stove.

The second case happens when forgetting switching off the gas stove while boiling some food. If the water of the food is vaporized by boiling, it can lead to a fire. Nowadays, the ratio of families that consists of one person or only old people is increasing and thus the probability of the second case fire is also increasing. Thus, we need to prevent the fire caused by forgetting switching off the gas stove, which is a main goal of the work.

* Corresponding author.

In order to prevent the second case fire, we use gas sensors. When the water is vaporized by continuous heat and the pots starts to be burnt, noxious gases such as carbon monoxide are generated. Thus, we can detect the early sign of the fire by detecting the noxious gases with the gas sensors. If the noxious gases are detected, the gas valve is automatically closed by the motor to turn off the gas stove and to prevent a serious fire. Then, the scheme notifies the generation of the noxious gases by making a phone call to the pre-stored number. Thus, the owner can notice the potential fire and deal with it.

The rest of the paper is organized as follows. We explain the design and the implementation of the presented work in section 2. The correct operation of the presented work is evaluated in section 3. Finally, we draw a conclusion in section 4.

2 Fire Prevention with Gas Sensors

The presented appliance supports four main functions. First, it detects the poisonous gases generated by the burn of the pots. Second, it detects the gas leak from the gas stove or the gas horse. Third, when detecting the poisonous gases or the gas leak, it automatically closes the gas valve. Fourth, when detecting the poisonous gases or the gas leak, the alarm system operates. The buzzer is sounded and a phone call is made to a pre-stored number. Thus, the presented appliance mainly consists of gas sensors to detect the poisonous gases and the gas leak, a motor to close and open the gas valve, DTMF dialer to make a phone call, buzzer, and a processor to control the components.

Fig. 1 shows the main hardware circuit of the appliance. The basic voltage is 12V, and the motor and the buzzer operate with 12V. Other components use 5V, and thus we use regulator to adjust the voltage. In order to prevent a back flow of current caused by the difference of the voltage, we deploy diodes.

There are two gas sensors. The gas sensor 1 detects the gas leak. It can detect hydrocarbon gases such as propane, methane, butane, and gasoline, ether, ester, and so on. The gas sensor 2 detects the burn of the food or the pots. It detects carbon monoxide and does not detect the vapor. Thus, boiling water does not cause the alarm.

The mechanism of detecting gases is as follows. When turning on the power supply, the temperature of the gas sensors is steeply increasing and the electric resistance of the gas sensors is rapidly decreasing. However, if the gas leak or the poisonous gases are detected by the sensor, the detected gas is adhered to the sensor, and the electric resistance of the sensor increases. Thus, we can know this event with the comparators. If the event is detected, it operates the motor to close the gas valve and initiates the alarm system. Fig. 2 depicts the entire flow chart.

The telephone number is stored using keypad. Fig. 3 shows the circuit diagram of AVR processor. The keypad is connected to the processor through port C. The telephone number that is typed with keypad is stored in EEPROM of the processor.

When detecting the gases, the processor outputs the stored telephone number to DTMF dialer through port A. DTMF dialer transforms the phone number to DTMF signal. Fig. 4 shows the circuit diagram of DTMF.

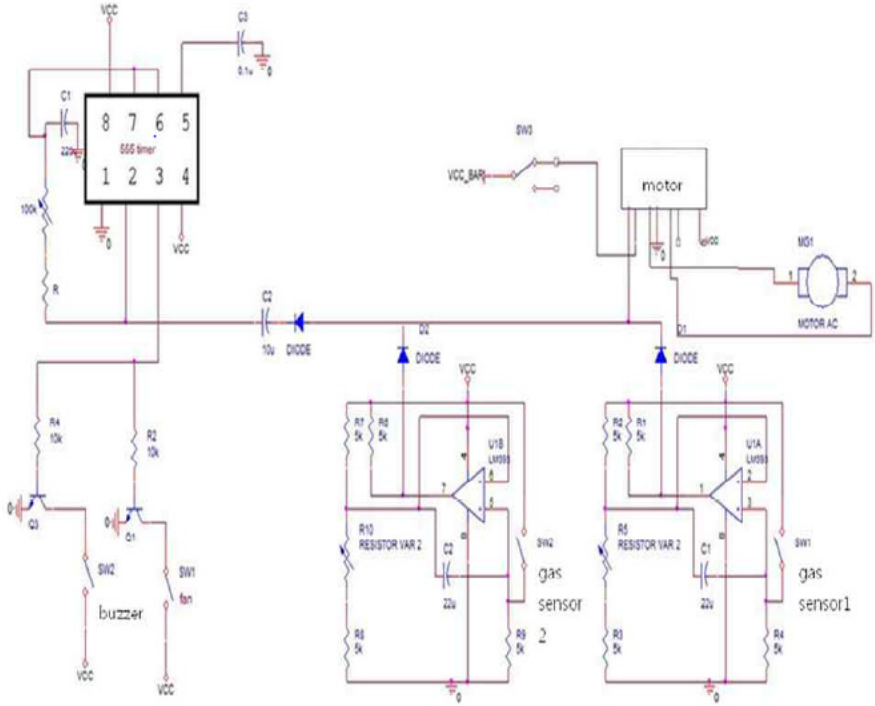


Fig. 1. Hardware Circuit of Presented Appliance

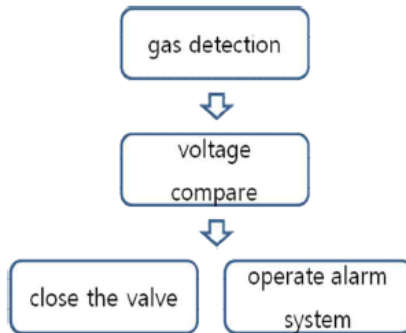


Fig. 2. Flow Chart of Presented Appliance

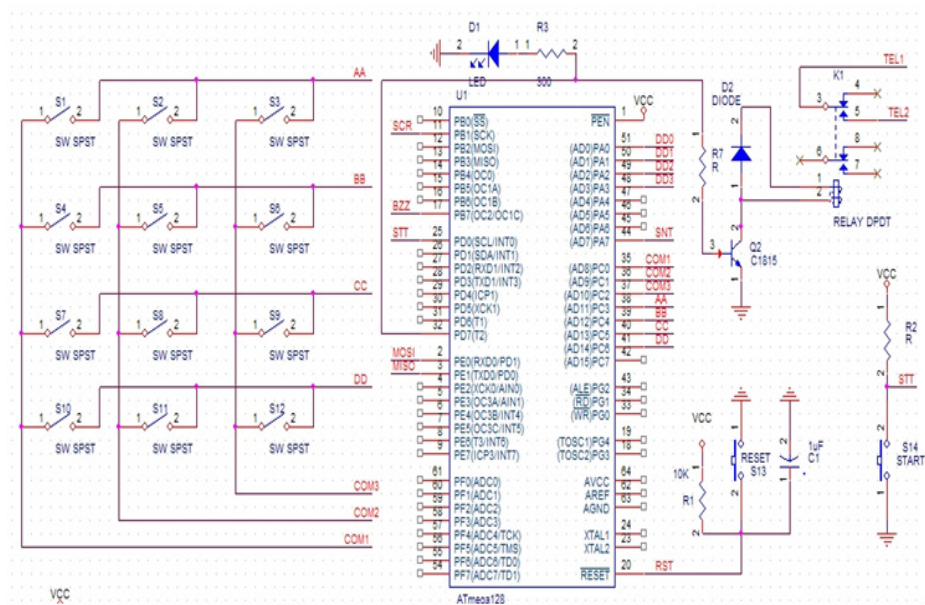


Fig. 3. Circuit Diagram of AVR Processor

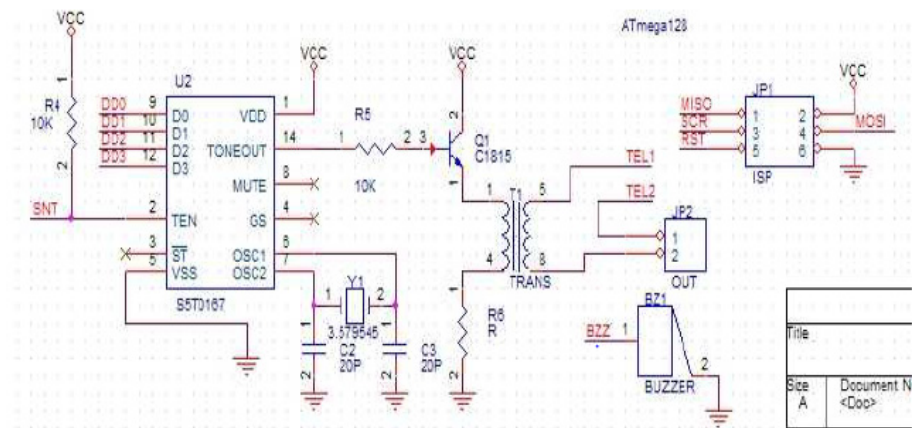


Fig. 4. Circuit Diagram of DTMF Dialer

3 Evaluation

We implement the prototype appliance on a bread board. The gas sensors are connected to comparators. The keypad to input the telephone number is connected to the AVR processor. We use SP-AQ2 sensor to detect the poisonous gases and KGS O2 sensor to detect the gas leak. Before we implement the prototype, we tested the target sensors with the circuit of fig. 5. In order to know the change of the electric

resistance of SP-AP2 sensor, we burned a paper and measured the electric resistance, which was increased from 9.7 K Ω to 95.8 K Ω . The change of the electric resistance of KGS 102 sensor was measured using a disposable lighter, which was increased from 23.3K Ω to 102.4 K Ω .

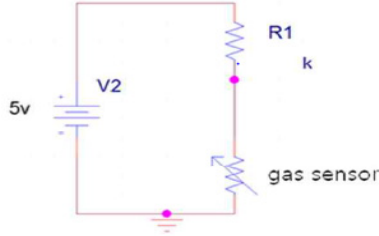


Fig. 5. Test Circuit of Gas Sensors

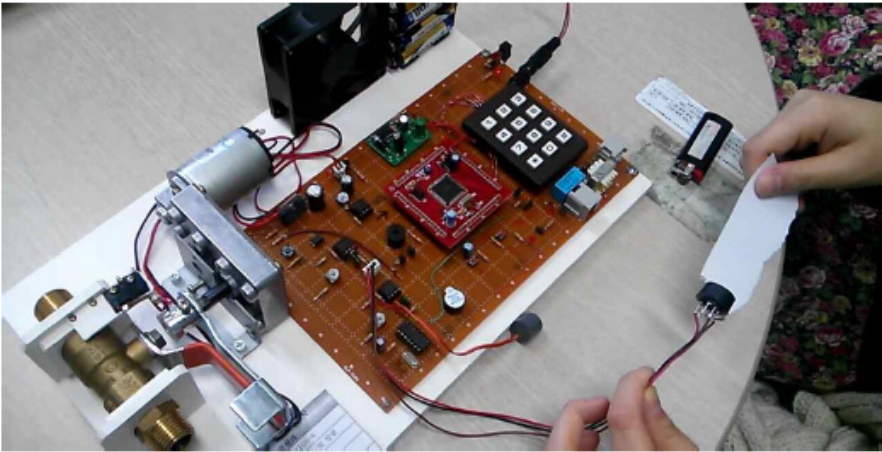


Fig. 6. Test of SP-AQ2 sensor

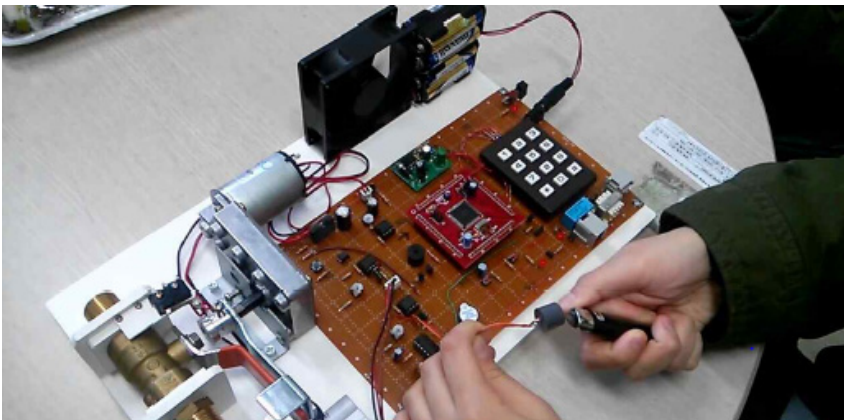


Fig. 7. Test of KGS 102 sensor

Figs. 6 and 7 show the test of each gas sensor in the prototype appliance. When burning a paper in front of the SP-AQ2 sensor, the poisonous gas was detected, which closed the gas valve using the motor and made a phone call to the pre-stored phone number (Fig. 6). Similarly, when emulating the gas leak with the disposable lighter, it was detected. The gas valve was closed and the phone call was made (Fig. 7).

4 Conclusion

In this work, we presented to use gas sensors in order to prevent a serious fire in the gas stove. The presented appliance detected the poisonous gases and the gas leak using sensors and closed the gas valve automatically on the detection of the emergent event. Users could be alarmed the emergent event with the telephone by storing the number beforehand.

The prototype appliance implemented on the bread board worked correctly. Burning a paper and the gas leak of the disposable lighter were successfully detected. The gas valve was automatically closed by the motor, and the phone call was made to the pre-stored number.

Acknowledgments. This study was financially supported by Seoul National University of Science and Technology

References

1. Baig, F., Beg, S., Khan, M.F.: Zigbee Based Home Appliances Controlling Through Spoken Commands Using Handheld Devices. *International Journal of Smart Home* 7(1), 19–26 (2013)
2. Chen, M., Wu, C.: A ZigBee-Based Home Control System Using OSGi Management Platform. *International Journal of Smart Home* 6(4), 15–28 (2012)
3. Ali, F., Kodirov, E., Ardiansyah, C.D., Lee, G.: Hey Home, Open Your Door, I'm Back! Authentication System using Ear Biometrics for Smart Home. *International Journal of Smart Home* 7(1), 173–182 (2013)
4. Chae, S.: Control Device for Cutting off GAS And Putting Outlights Automatically. Korea Patent (2), 003,402,710,000 (2004)

Word Spotting Application in Historical Mongolian Document Images

Hongxi Wei and Guanglai Gao

School of Computer Science, Inner Mongolia University, Hohhot, China
{cswhx, csggl}@imu.edu.cn

Abstract. This paper proposes a framework based on the word spotting technology for indexing and retrieving the historical Mongolian document images. In the framework, the scanned document images are segmented into word images by some preprocessing steps such as binarization, connected component analysis and so on. And then each word image is processed by the following procedure, including removing inflectional suffixes, feature extraction and fixed-length representation. Finally, each word image is represented by a fixed-length feature vector and considered as an indexing term. At the retrieval stage, the necessary query keyword image can be obtained by synthesizing a sequence of glyphs according to the spelling rules of Mongolian language. For word matching, the query keyword image is also converted into a fixed-length feature vector through the same procedure. And a ranking list can be returned in descending order of similarities between the query keyword image and each candidate word image. Experimental results on the data set prove the feasibility and effectiveness of the proposed framework.

Keywords: Word spotting, document image retrieval, feature extraction, query keyword image.

1 Introduction

A large amount of historical Mongolian documents are collected in many libraries of China, especially in Inner Mongolia Autonomous Region. One of the most famous is the Mongolian Kanjur, which is a Mongolian encyclopedia including religion, history, medicine, astronomy and so on. The Mongolian Kanjur was made by woodblock printing in 1720 (Qing Dynasty). This edition of the Mongolian Kanjur contains 108 volumes in total, or about 45,000 pages with twenty million words more or less. Only eight sets survive in all over the world and one of them is collected in the library of Inner Mongolia University. To protect the invaluable Mongolian Kanjur as long as possible, librarians of Inner Mongolia University have scanned each page into a digital image. Thus, the public can browse these digital images more convenient and faster such as via Internet. However, it is difficult to retrieve them due to a lack of indexing.

In the field of document image retrieval (DIR), optical character recognition (OCR) is the most common way to create the indexing. However, the words in the Mongolian Kanjur are equivalent to a kind of off-line handwritten words and the

scanned images are always with high degradations because of the passage of time. Hence, it is more difficult to segment this kind of off-line handwritten words into correct characters.

In [1], Gao et al. presented a segmentation-based method to recognize the words of the Mongolian Kanjur. On a data set of 5,500-word images with a good quality, the recognition rate of the words is only 71%. For a low image quality, the recognition rate will be even lower. Therefore, the OCR technology is still challenging for recognizing the words of the Mongolian Kanjur.

In the field of DIR, word spotting technology is an alternative to OCR. It is a content-based retrieval procedure that spots words directly on document images without the OCR procedures [2, 3]. So, the word spotting technology is especially suited to the historical document image retrieval. Word spotting was firstly introduced into DIR by Manmatha et al. [4]. After that, word spotting was widely used for indexing and retrieving historical document images in many different languages, such as English [5], Greek [6], Japanese [7], Ottoman [8] and so on. As far as we know, the word spotting technology has not yet been used for the retrieval of the historical Mongolian document images.

In this paper, we propose and implement a framework of word spotting for the Mongolian Kanjur images. Therein, each word image of the Mongolian Kanjur is converted into a feature vector with the fixed length. Thus, the similarity between word images can be measured by calculating the Euclidean distance on their feature vectors. Moreover, the morphological variations on word images can be overcome by removing the inflectional suffixes from word images. Consequently, not only the same form but also the inflectional forms of the query keyword can be obtained during the retrieval. Additionally, query keyword images can be generated by synthesizing a sequence of glyphs at the retrieval stage.

The rest of this paper is organized as follows. Section II describes the proposed framework in detail. Experimental results are shown in Section III. Finally, Section IV gives conclusions and sheds lights on the future works.

2 Proposed Framework

The main steps of word spotting are presented in [3]. Firstly, the document image is segmented into words. And then the suitable features are extracted from the words. Finally, the matching and ranking are performed on the feature vectors of the words. Based on the above steps, a word spotting framework for the Mongolian Kanjur images is proposed and the whole structure is depicted in Fig. 1. In the framework, all the scanned Kanjur images should be segmented into the corresponding word images for the further processing. In our previous work [9], an effective preprocessing approach has been proposed to obtain the words from the scanned Kanjur images. Hence, this paper only concentrates on the steps after word segmentation, including removing inflectional suffixes, feature extraction, the fixed-length representation and so on. The details are described in the following subsections.

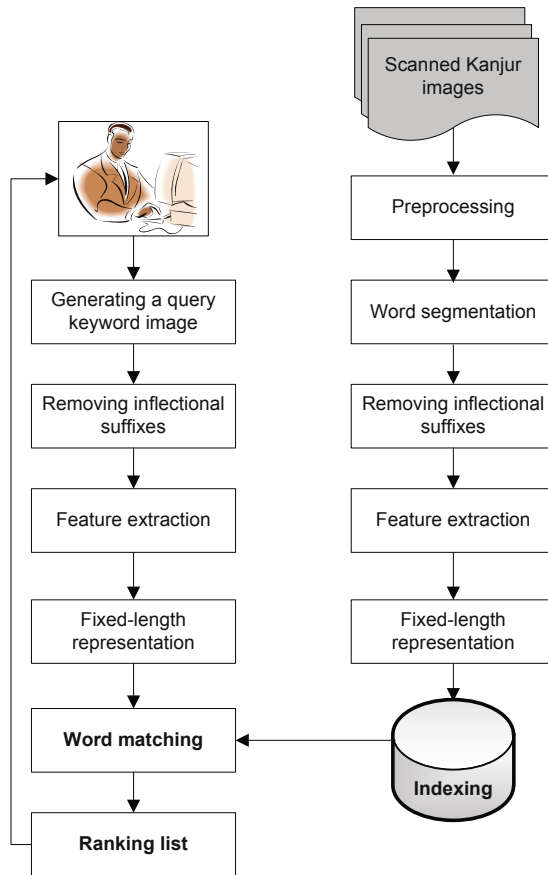


Fig. 1. The whole structure of the proposed framework

2.1 Removing Inflectional Suffixes

The Mongolian language is a kind of agglutinative language. Its word formation and inflection is built through connecting different suffixes to the roots or stems. The suffixes are classified into two categories. One is word-formation suffix that can produce variations of part-of-speech or meaning. The other one is word-inflection suffix that often causes variations of person or tense.

Therefore, there are lots of words with the same part-of-speech and meaning, but they contain different inflectional suffixes at the end of the words. Once the inflectional suffixes are removed from the word images, not only the same form but also the inflectional forms of the query keyword in the other person or tense can be returned in the ranking list. To attain the purpose, an approach is shown in [10] to determine whether or not one word image contains inflectional suffix. If a word image

contains some kind of inflectional suffix, the inflectional suffix would be segmented from the word image and the rest part or else the whole word is used for feature extraction.

2.2 Feature Extraction

Profile-based features are the most popular and widely used to represent historical handwritten word images [5], which can capture the external shape and the internal structure of word images. So, four profile features, including *horizontal projection profile*, *left profile*, *right profile* and *horizontal background-to-foreground transitions*, are extracted from per image row. Another four profile features, including *vertical projection profile*, *upper profile*, *lower profile* and *vertical background-to-foreground transitions*, are extracted from per image column. As a result, every word image in the Mongolian Kanjur as well as the query keyword image are all represented by the eight profile feature vectors.

2.3 Fixed-Length Representation

Due to varying sizes of word images, the lengths of the profile feature vectors are different from each other. Although dynamic time warping (DTW) algorithm can partly handle the varying length of the profile feature vectors [11], the computational cost is so expensive that DTW is not competent for the task of the on-line image matching at the retrieval phase.

Table 1. The methods of the fixed-length representation

Fixed-length representation method	Description
<i>DFT coefficients representation</i>	For each kind of profile feature, DFT is performed on the profile feature vector and a number of lower-order complex coefficients are used to approximately represent the original profile feature vector.
<i>DFT moduli representation 1</i>	For each kind of profile feature, DFT is performed on the profile feature vector and a number of moduli of lower-order complex coefficients are used to approximately represent the original profile feature vector.
<i>DFT moduli representation 2</i>	For each kind of profile feature, DFT is performed on the profile feature vector and a number of moduli of lower-order complex coefficients (<i>without the modulus of the first complex coefficient</i>) are used to approximately represent the original profile feature vector.
<i>DCT coefficients representation</i>	For each kind of profile feature, DCT is performed on the profile feature vector and a number of lower-order coefficients are used to approximately represent the original profile feature vector.

To improve the speed of word matching, we have presented four methods (see Table 1) to convert the profile feature vectors with varying length into the fixed-length feature vectors [12]. Fig. 2 shows the comparison results of the four methods (as in [12]). The method of "DFT moduli representation 2" is the best one and the appropriate number of moduli is ten for each profile feature. In this way, every word image can be converted into a fixed-length feature vector with eighty values (i.e. eighty moduli). Thus, the process of word matching can be realized by calculating the similarity (e.g. Euclidean distance) between the query keyword and the candidate word on their fixed-length feature vectors. And a ranking list can be returned in descending order of the similarities.

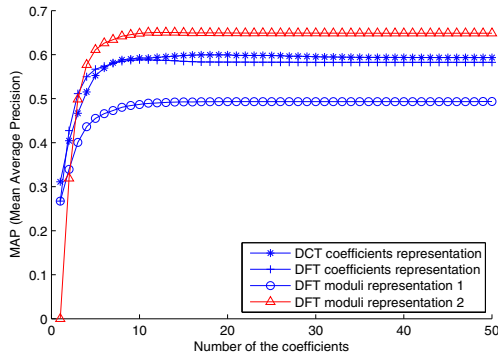


Fig. 2. The comparison results of the four fix-length representation methods

2.4 Synthesizing Query Keyword Images

In order to provide query keyword images for retrieving, we proposed an approach to synthesizing Mongolian word images by splicing a number of glyphs according to the spelling rules of Mongolian language. Mongolian language is a kind of cursive alphabetic script. All letters of one Mongolian word should be conglutinated together from top to bottom along with the writing direction. Based on the characteristics, a set of glyphs was determined in advance, which can generate any Mongolian word image.

In the process of synthesizing a word image, the following two aspects must be guaranteed: (1) the neighboring glyphs must be aligned at their baselines along with the writing direction; (2) the neighboring glyphs must be seamlessly connected together at their baselines. Fig. 3 describes the procedure of splicing glyphs. In this way, any desired query keyword image can be formed.

2.5 Pruning in Word Matching

In word matching phase, the pruning can be used to improve the retrieval efficiency. In the proposed framework, a pruning strategy has been adopted by comparing some scalar values (i.e. area and aspect ratio) of word images before calculating similarities. The following two criteria should be satisfied simultaneously for pruning:

$$\frac{1}{\alpha} \leq \frac{Area_{query}}{Area_{candidate}} \leq \alpha, \alpha=1.7$$

$$\frac{1}{\beta} \leq \frac{Aspect_{query}}{Aspect_{candidate}} \leq \beta, \beta=1.4 \quad (1)$$

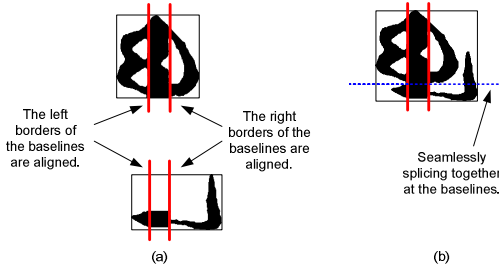


Fig. 3. The procedure of splicing glyphs

3 Experimental Results

To evaluate the performance of the proposed framework, 200 pages scanned images were selected randomly from the collection of the Mongolian Kanjur. Each page was segmented into the corresponding word images by our previous approach in [9]. As a result, there are 47106 word images in total to form the experimental data set. And each word image has been annotated using Unicode.

By analyzing the annotations, forty meaningful words with the number of instances more than 20 were determined. They were taken as the query keywords (see Table 2) and used to evaluate the performance of the proposed framework. Evaluation metrics are *MAP* (mean average precision), *R-Precision* and *Precision@k*.

3.1 Experiment I

In this experiment, we evaluated the performance of the fixed-length representation. Each word from the Mongolian Kanjur was converted into the fixed-length feature vector without removing inflectional suffixes. At the retrieval stage, each query keyword was synthesized and then was also converted into the corresponding fixed-length feature vector. For each query keyword, the average precision (AP), R-Precision and Precision@k can be calculated according to the ground truth.

The average precisions and R-Precisions are shown in Fig. 4 and Fig. 5. The MAP and average R-Precision are **49.97%** and **50%** separately. The Precision@k of each query keyword is listed in Table 3. Here, the values of the parameter k are 1, 5, 10, 15 and 20. From Table 3, we can see that the performance of the top candidates in the returned results are confirmed.

Table 2. The list of query keywords

Sequence No.	1	2	3	4	5	6	7	8	9	10
Keyword	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ
Instances	296	69	218	68	55	93	128	111	64	111
Sequence No.	11	12	13	14	15	16	17	18	19	20
Keyword	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ
Instances	154	33	38	116	21	439	222	35	41	136
Sequence No.	21	22	23	24	25	26	27	28	29	30
Keyword	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ
Instances	99	76	32	187	49	74	30	76	142	498
Sequence No.	31	32	33	34	35	36	37	38	39	40
Keyword	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ	ᠰᠠᠨᠠ
Instances	182	105	32	311	42	103	83	293	183	82

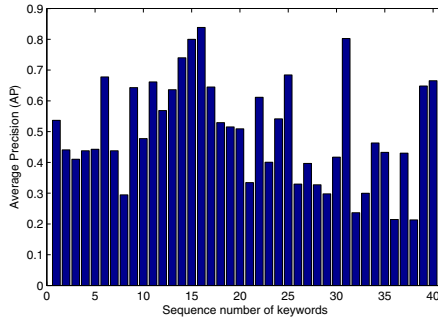


Fig. 4. The average precisions of the 40 query keywords

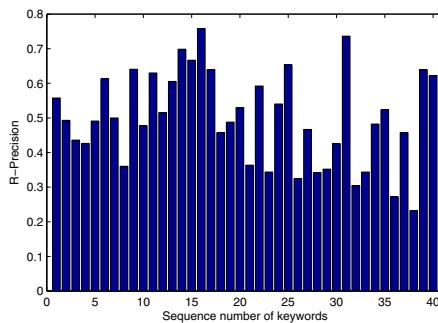


Fig. 5. The R-Precisions of the 40 query keywords

Table 3. The Precision@k of each query keyword

Keyword	Precision @1	Precision @5	Precision @10	Precision @15	Precision@20
1	100.00%	100.00%	100.00%	93.33%	85.00%
2	100.00%	100.00%	90.00%	80.00%	85.00%
3	100.00%	100.00%	100.00%	93.33%	95.00%
4	100.00%	100.00%	90.00%	86.67%	70.00%
5	100.00%	100.00%	90.00%	80.00%	70.00%
6	100.00%	100.00%	100.00%	100.00%	100.00%
7	100.00%	100.00%	80.00%	80.00%	85.00%
8	100.00%	100.00%	90.00%	93.33%	90.00%
9	100.00%	100.00%	90.00%	93.33%	95.00%
10	100.00%	100.00%	90.00%	80.00%	85.00%
11	100.00%	100.00%	100.00%	100.00%	100.00%
12	100.00%	100.00%	90.00%	80.00%	70.00%
13	100.00%	100.00%	90.00%	86.67%	80.00%
14	100.00%	100.00%	100.00%	100.00%	100.00%
15	100.00%	100.00%	100.00%	80.00%	70.00%
16	100.00%	100.00%	100.00%	100.00%	100.00%
17	100.00%	100.00%	100.00%	93.33%	95.00%
18	100.00%	100.00%	100.00%	93.33%	75.00%
19	100.00%	100.00%	90.00%	86.67%	80.00%
20	100.00%	100.00%	100.00%	93.33%	90.00%
21	100.00%	100.00%	100.00%	86.67%	85.00%
22	100.00%	100.00%	100.00%	100.00%	100.00%
23	100.00%	100.00%	70.00%	46.67%	45.00%
24	100.00%	100.00%	100.00%	86.67%	75.00%
25	100.00%	100.00%	90.00%	86.67%	85.00%
26	100.00%	100.00%	90.00%	80.00%	70.00%
27	100.00%	100.00%	90.00%	80.00%	80.00%
28	100.00%	100.00%	80.00%	73.33%	65.00%
29	100.00%	100.00%	100.00%	93.33%	85.00%
30	100.00%	100.00%	100.00%	100.00%	100.00%
31	100.00%	100.00%	100.00%	100.00%	100.00%
32	100.00%	100.00%	80.00%	60.00%	60.00%
33	100.00%	100.00%	70.00%	53.33%	45.00%
34	100.00%	100.00%	100.00%	93.33%	85.00%
35	100.00%	100.00%	90.00%	80.00%	75.00%
36	100.00%	80.00%	70.00%	66.67%	60.00%
37	100.00%	100.00%	100.00%	100.00%	95.00%
38	100.00%	100.00%	100.00%	93.33%	85.00%
39	100.00%	100.00%	100.00%	86.67%	90.00%
40	100.00%	100.00%	100.00%	100.00%	95.00%
Average	100.00%	100.00%	93.00%	86.00%	83.00%

3.2 Experiment II

In this experiment, we concentrated on the effectiveness of the pruning strategy in the word matching phase. In the same way as *Experiment I*, the word images and the query keyword images were directly converted into the fixed-length feature vectors without removing inflectional suffixes.

The experimental results are given in Table 4, which proves that the pruning strategy can improve the retrieval efficiency. The average retrieval time is shortened to one fifth at the cost of a few loss of the performance.

Table 4. The effectiveness of the proposed pruning scheme

Scheme	MAP	Avg R-Precision	Avg retrieval time
<i>Pruning</i>	49.97%	50.00%	10.34s
<i>No Pruning</i>	50.51%	50.69%	50.83s

3.3 Experiment III

In this experiment, the retrieval performance after removing inflectional suffixes was evaluated. To reach the goal, the inflectional suffixes were removed from word images which contains some kind of inflectional suffix. Then, the whole word or the rest part was converted into the corresponding fixed-length feature vector. Before being retrieved, the inflectional suffixes were also removed from the query keyword images. In Table 2, there are 14 query keywords with inflectional suffixes. The sequence numbers are 2, 9, 10, 13, 15, 20, 24, 25, 27, 28, 31, 34, 39 and 40. They are used for evaluating the retrieval performance after removing inflectional suffixes. The comparison performance between retaining and removing inflectional suffixes is shown in Table 5. We can see that the morphological variations of word images can be overcome by removing the inflectional suffixes. Moreover, not only the same form but also the inflectional forms of the query keyword in the other person or tense can be returned in the ranking list.

Table 5. The comparison performance between retaining and removing inflectional suffixes

Scheme	MAP	Avg R-Precision
<i>Retaining inflectional suffixes</i>	50.79%	51.59%
<i>Removing inflectional suffixes</i>	54.82%	55.32%

4 Conclusions

This paper proposes an efficient framework for spotting words on the Mongolian Kanjur images. On our experimental data set, the MAP and average R-Precision are **54.82%** and **55.32%** respectively using the pruning strategy and removing the inflectional suffixes, which proves the proposed framework can be competent for the task

of historical Mongolian document images retrieval. However, several word images obtained from the Mongolian Kanjur were degraded because of aging. So, our future work will be focused on the methods and technologies for improving the image quality. Additionally, the relevant feedback technique or the pseudo relevant feedback technique can improve the ranking list in general. Hence, integrating this aspect into the proposed framework is also the future work of our research.

Acknowledgments. This work is supported by Natural Science Foundation of China (Grant No.60865003). This work is also supported by Natural Science Foundation of Inner Mongolia Autonomous Region and the project numbers are 2012MS0921 and 2011ZD11.

References

1. Gao, G., Su, X., Wei, H., Gong, Y.: Classical Mongolian Words Recognition in Historical Document. In: Proceedings of ICDAR 2011, pp. 692–697 (2011)
2. Louloudis, G., Kesidis, A.L., Gatos, B.: Efficient Word Retrieval Using a Multiple Ranking Combination Scheme. In: Proceedings of DAS 2012, pp. 379–383 (2012)
3. Kesidis, A.L., Gatos, B.: Efficient Cut-off Threshold Estimation for Word Spotting Applications. In: Proceedings of ICDAR 2011, pp. 279–283 (2011)
4. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing Handwriting Using Word Matching. In: Proceedings of the ICDL 1996, pp. 151–159 (1996)
5. Rath, T.M., Manmatha, R.: Features for Word Spotting in Historical Manuscripts. In: Proceedings of ICDAR 2003, pp. 218–222 (2003)
6. Gatos, B., Konidakis, T., Ntzios, K., Pratikakis, I., Perantonis, S.J.: A Segmentation-free Approach for Keyword Search in Historical Typewritten Documents. In: Proceedings of ICDAR 2005, pp. 54–58 (2005)
7. Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace Method for Text Retrieval in Historical Document Images. In: Proceedings of ICDAR 2005, pp. 437–441 (2005)
8. Ataer, E., Duygulu, P.: Matching Ottoman words: An Image Retrieval Approach to Historical Document Indexing. In: Proceedings of CIVR 2007, pp. 341–347 (2007)
9. Wei, H., Gao, G., Bao, Y., Wang, Y.: An Efficient Binarization Method for Ancient Mongolian Document Images. In: Proceedings of ICACTE 2010, pp. 43–46 (2010)
10. Wei, H., Gao, G., Bao, Y.: A Method for Removing Inflectional Suffixes in Word Spotting of Mongolian Kanjur. In: Proceedings of ICDAR 2011, pp. 88–92 (2011)
11. Rath, T.M., Manmatha, R.: Word Image Matching Using Dynamic Time Warping. In: Proceedings of CVPR 2003, pp. 521–527 (2003)
12. Wei, H., Gao, G., Zhang, X.: Indexing for Mongolian Kanjur Images in Word Spotting. *Journal of Computational Information Systems* 9(4), 1501–1508 (2013)

Binary Program Statistical Features Hiding through Huffman Obfuscated Coding

Xiaopeng Niu, Qingbao Li, Wei Wang, and Xiaokang Weng

National Digital Switching System Engineering and Technological Research Center
Zhengzhou, China
xiaoyuer8082@163.com

Abstract. Mutants produced by current metamorphic engine are divers, but they still contain shortcomings that reliably distinguish them from normal program. This paper introduces a novel binary obfuscation technique with the potential of evading both statistical and semantic detections. It transforms the binary program into mimicry executables that exhibit high similarity to benign programs in terms of statistical properties and semantic characteristics. Experimental results show that the mimicry executables are indistinguishable from benign programs in byte frequency distribution and entropy, and no false instructions produced.

Keywords: binary program, obfuscate, Huffman, digesting, encoding, decoding.

1 Introduction

For special program that need protection, how to hide the real behavior of it is a concrete problem for research, and that is the aim of this paper. Recent studies demonstrate that malware authors need concern themselves with the distribution of their creations to end user systems no longer, and they can leave the task to specialized pay-per-install services [1]. In such an environment, the primary concern of malware is hiding the feature of the code and evading detection as it carries its task.

Although the classic polymorphism and metamorphism enable software to generate many binary instances with different byte patterns, they cannot effectively disguise the presence of malicious code in terms of statistical properties and program semantics [2]. Compression and encryption in polymorphism usually change the statistical characteristics of a program in such a dramatic manner that the protected program can be easily classified as suspicious and be further scrutinized [3]. Vishwath [4] proposed a prototype system, Frankenstein, that stitching malware from benign binaries to hide distinguishing semantic features, but the method is still not mature.

Wu Zhenyu [5] developed an approach for binary code obfuscation, called mimic-morphism, which transformed an executable binary into a different one that with statistical and semantic characteristics highly similar to those of the mimicry target. But the Key data structure “CommomInst” and the core phase “Digesting” have some shortcomings that can lead to false instruction encoding. For the structure of

“CommonInst”, the method treats parameters equally, without any classification and pretreatment dealing with each parameter in different locations. And if some false instructions appear in the encoding phase, there is no reparation to deal with the fault instruction. Based on the work of Wu Zhenyu [5], this paper develops a new approach that hiding the statistical characteristic and the behavior information of program, called refined-mimimorphism. The main contributions of this paper that distinguishes the work of former researcher are the method of classifying the parameter in different location and the method of correcting the fault instruction in encoding phase, which can improve the correctness of instruction encoding.

2 Design Principle

The main process of refined-mimimorphism is character mimic transformation, which changes the given input data into certain output data with a different type of statistical properties, thereby concealing the true identity of the original data. The idea of mimic function was first introduced by Peter Wayner [6], which was the functional inverse of the Huffman coding and mainly used for encryption. Here the refined-mimimorphism contains three major parts: digesting function, encoding and decoding function.

The different components of refined-mimimorphism are diagrammatically represented in Figure 1. Digesting function extracts the statistical specialty of the target binary code and organizes it as a Huffman tree. Encoding function transforms the binary code to be obfuscated based on the tree, and decoding function uncovers the input data from the mimicry output.

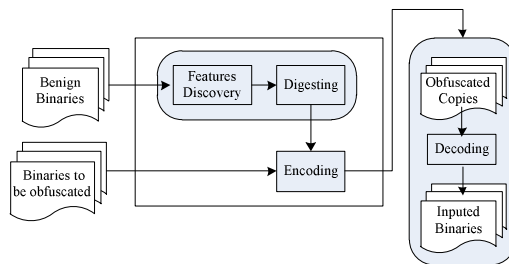


Fig. 1. High-Level Architecture of Refined-Mimimorphism

3 Design Method

Refined-mimimorphism trains the mimicry target and extract the statistical features of the target, record the features as template for obfuscated coding. Below, we describe the detailed method of it, which differs from those of [5].

3.1 Digesting

Preparing for the digesting function, the disassemble module translates the instructions in the executable binary streams into assembly codes and store them in database.

The digest function records the instruction information in database with Refined-CommonInst structures, as shown in Figure 2.

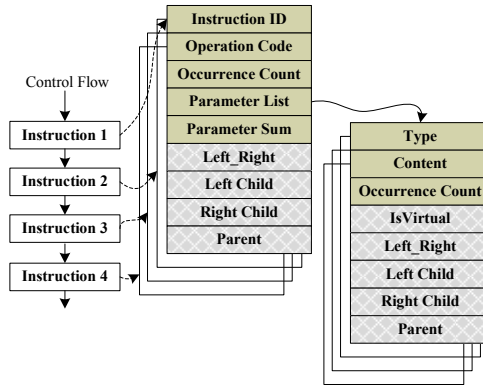


Fig. 2. RefinedCommonInst Structure

Algorithm 1 Degesting Target Code

```

dasm: Assembled instruction sequence
tmpInst: RefinedCommonInststructure
asmVector: Current IDV
while dasm is not empty do
  instCount = instCount +1;
  instID = CheckCodeIsIn(opCode, asmVector);
  If (instID>= 0)
    Modify the information in asmVector;
  Else
    InitializetmpInst;
    Append tmpInst to asmVector;End if
  If (instCount == Tree_order)
    instPfxId++;
    BuildHufTreeOfParas (instPfxId, asmVector);
    for(int i=0; i<asmVector.size(); i++){
      TreeWalkOfParas (instPfxID, asmVector);
    }
    BuildHufTreeOfInst (asmVector);
    TreeWalkOfInst (instPfxID, asmVector);End if
  Dasm->MoveNext ();
end while

```

After disassembling, the digesting function processes the decoded instructions in a sequential manner. Internal to the digest function, a sequence of most recently processed instruction IDs, called InstPfx, is maintained, acting as the “symbol prefix” of the mimic function. For each instruction, the digest function first tries to locate an instruction digest vector (IDV) associated with the current InstPfx. If absent or the

vector is filled, a new vector is created. Then the digest function reconstructs the information of the instruction into RefinedCommonInst structure. Each IDV consists of parameter detail records (PDRs), indexed by the instruction ID.

At the end of the digesting phase, each IDV is converted into an instruction Huffman tree (IHT), based on the frequency counter of each instruction. Correspondingly, each PDR is turned into an instruction encoding template (IET) by converting all the frequency tables associated with the prefix and parameters into Huffman trees.

After the digesting process, instructions in InstPfx are reconstructed as IHT and IET. Algorithm 1 provides a high level overview of the mimimorphic digesting operations. Because different instructions have different number of parameters, some has one, while others have two or three. Parameters in different locations have different addressing mode, representing a kind of instruction characteristic. In order to mimic this property, so we distinguish parameters in different locations, constructing IET for the parameter in each location.

3.2 Encoding

Utilizing the digesting result (IHTs and IETs), the encoding phase transforms an arbitrary piece of binary (P) into a sequence of executable (P') instructions that resembles the mimicry target (PP), as shown in formula 1.

$$F(P) = P' \approx PP \tag{1}$$

The mimicry executables P' hide most of the statistical information of P. Three components of the encoding process including the PRNG, the encoding function, and the assembler are shown in figure 3:

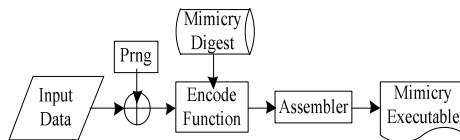


Fig. 3. Encoding Process [5]

Algorithm 2 provides an overview of the encode function in figure 3. The IHTs generated in digesting process are stored in XML file and ordered with InstPfxId. Utilize InstPfxId sequentially can locate an IHT, and searches for an IET from the IHT by travelling the Huffman tree from the tree node with pre-order, taking left or right branches according to the randomized data in SBin. SBin is generated by randomize function, the input data Bin is randomized by XORing with a pseudo-random data stream. The randomization can generate random data that satisfy the mimic function (uniformly distributed), and erase the characteristics of the original binary. Similar to locating IET in IHT, the parameter of instruction is located in IET by travelling it. Using the operate code and parameter in IET, the instruction can be constructed easily.

Because parameters of the instruction are constructed according to the randomized data, so the parameters may not fit the location of instruction. For example, the MOV instruction not permits data transformation from memory space directly to memory space, and between the same registers in 51 instruction system. Instructions “MOV 11H, 14H” and “MOV A, A”, that generated by binary data 1001 are false instructions violating coding principle, as shown in figure 4.

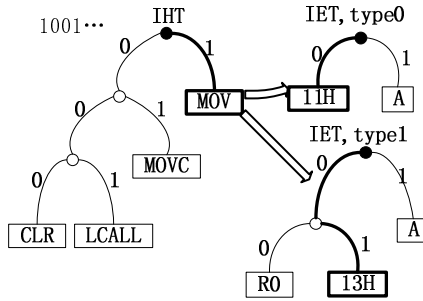


Fig. 4. A case of problematical instruction

The function CheckPara(para) in while loop is to check the validity of parameters that generated by TreeWalkIet(IET,SBin). If errors occur, change the randomized data SBin randomly, and walk the parameter template IET again until got the legal parameters.

Algorithm 2 Encoding Function

```

Bin: Input binary data
Digest: Mimicry digest
RSeed: Pseudo-RNG seed
InstPfxId =1; isValid = false;
SBin = Randomize(Bin,RSeed);
IHT=FindIht(InstPfxId);
While SBin is not empty do
  if(IHT has not been over travelled){
    IET = TreeWalkIht(IHT,SBin);
    While(!isValid){
      para = TreeWalkIet(IET,SBin);
      isValid = CheckPara(para);}
    Inst = Construct(IET,para);
    Append Inst to InstCollection;
    Update SBin;}
  Else{
    InstPfxId++;
    IHT = FindIht(InstPfxId);}
  end while
Result =Assemble(InstCollection);

```

3.3 Decoding

The decoding phase is the inverse of the encoding phase. Based on the trees generated in digest phase, the decoding phase uncovers the input data from the mimicry output. Three components of the mimimorphic engine involved in this phase: the disassembler, the decoding function, and the PRNG component.

Algorithm 3 Decoding Function
Mimicry_{In}: Mimicry instance
Digest: Mimicry digest
RSeed: Pseudo-RNG seed
 Initialize *InstPfx*;
InstCollection = Disassemble(*Mimicry_{In}*);
For each *Inst* in *InstCollection* **do**
 IHT = Lookup(*Digest*, *InstPfx*);
 (*IET*, *IData*) = NodeLookup(*IHT*, *Inst*);
 IData = InstDecode(*IET*, *Inst*);
 Append *IData* to *Data_{Rand}*;
 Update *InstPfx* with *Inst*;
end for
 Result = DeRandomize(*Data_{Rand}*, *RSeed*);

The high level description of the mimimorphic decoding operation is given in Algorithm 3. The *InstPfx* is used to record the most recently decoded instruction IDs. A mimicry instance is first disassembled into *RefinedCommonInst* structures, before being processed sequentially. The *NodeLookup* function locates the *IET* in the *IHT* with the matching instruction ID. Meanwhile, it produces a stream of data bits that corresponds to the branches taken from the root of the Huffman tree to the leaf node. The *InstDecode* function further retrieves the data bits encoded in each mimicry instruction by performing Huffman encoding operations. Finally, the *DeRandomize* function uncovers the original data by XORing the decoded data with a pseudo-random data stream, which are generated by the PRNG with the same seed used in the encoding phase.

4 Analysis and Test

Finally, to check the efficiency and effectiveness of the approach proposed in this paper, three kinds of test are taken, including run time analysis, coding error rate test and statistical test.

4.1 Run-Time Efficiency

Let n denotes the number of possible binary symbols in a given program, R denotes the order of a mimic function, and C denotes the size of parameter vector. In the

digesting phase, collecting symbol usages and constructing the RefinedCommonInst take linear time. Then, converting all symbol frequency tables into Huffman trees takes sub-linear time, with a constant bound-the total number of input length. Overall the digesting phase runs in linear time. The encoding and decoding phases essentially consist of a prefix lookup followed by a Huffman decoding and encoding, which are constant time operations for each input or output symbol. Therefore, they are linear too. As shown in Table 1, the whole mimic function has a linear time computational complexity.

Table 1. Runtime analysis for mimic function

Digesting=O(n)		
CommonInstConstructing	$2*n*R+n*C+C*R*n$	
IHT and IET Building	$n*R$	
Input length	n	
EnCoding/DeCoding=O(n)		
For each symbol	Locate Huffman Tree	1
	Huffman de(en)coding	1
	Input length	n

Although both the mimimorphism in [5] and the refined-mimimorphism in this paper have a linear time complexity, for the reason that the refined-mimimorphism uses a more elaborated structure (RefinedCommonInst) and additional false instruction correcting operation is taken, it needs more time than mimimorphism. The slop of refined-mimimorphism is a little greater than that of mimimorphism, increased by about 10%, as shown in figure 5.

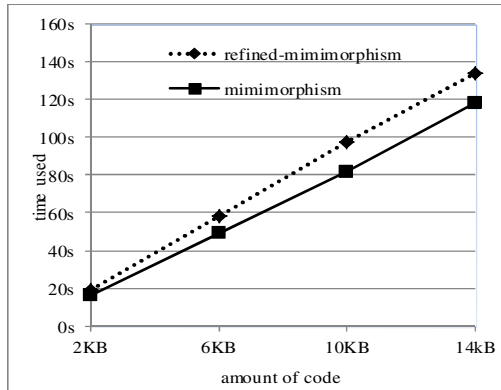


Fig. 5. Run time test for mimimorphism and reined-mimimorphism

4.2 Coding Error Rate Test

Because the method in [5] didn't classify the parameter type, so it could make several errors during the encoding phase, generating illegal instruction. The errors can be classified into four categories:(1) taking immediate value as destination operand, such

as MOV #08H, 15H, called error type 1; (2) taking indexed addressing as destination operand, such as MOV @A+DPTR, A, called error type 2; (3) invalid register using, such as CJNE 31H, #08H, A, called error type 3; (4) direct memory data moving, such as MOV 13h, 14h, called error type 4. The instruction encoding error rate of method in [5] is showed in figure 6.

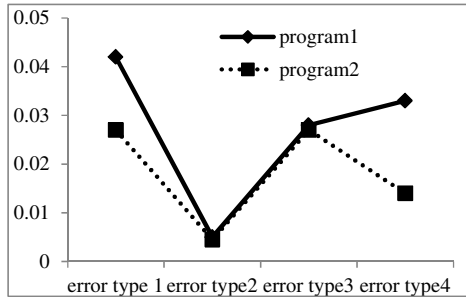


Fig. 6. Error rate of four error types for mimimorphism

As the method proposed in this paper classifies parameters according to the location of instruction, and revises parameters after constructing. So the encoding process will not produce the errors of the four types, the error rate is 0, better than that in [5]. Although the high exactness can affect much time efficiency, as shown in figure 6, they are both at one level time complexity, within the tolerance of application.

4.3 Statistical Test

We run the output of refined-mimimorphism against statistical tests, namely the Kolmogorov Smirnov (K-S) and byte entropy tests, the same as test method in [5]. The K-S test is a general purpose statistical test, whereas the byte entropy is proposed for detecting packed or encrypted malware [7].

The K-S test $kstest2(x_1, x_2)$ performs a two-sample Kolmogorov-Smirnov test to compare the distributions of the values in the two data vectors x_1 and x_2 . The null hypothesis is that x_1 and x_2 are from the same continuous distribution. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, 0 otherwise. And when the p -value is less than the predetermined significance level α , which is often 0.05, indicating that the observed result would be highly unlikely under the null hypothesis. The $kstest2$ is:

$$[h, p] = kstest2(x_1, x_2) = \max(|F_1(x) - F_2(x)|) \quad (2)$$

Where $F_1(x)$ is the proportion of x_1 values less than or equal to x and $F_2(x)$ is the proportion of x_2 values less than or equal to x . For the experiments, we perform K-S test between mimicry file and raw binary file, between mimicry file and legitimate file.

The byte entropy test is based on the randomness of compressed or encrypted files. The byte entropy test measures the randomness of the distribution of bytes:

$$entropy(X) = -\sum_x P(x) \log P(x) \tag{3}$$

Where X is a byte sample and P(x) is the probability P(X=x). For the experiments, we measure the byte entropy of different test samples, either mimicry or legitimate files. If the entropy of mimicry file is much higher than that of legitimate file, then mimicry process is a failure. The statistical test results are showed in table 2:

Table 2. Statistical Test Result

	Test file types	Mean	Std. dev.
K-S test	Raw binaries	0.091(p)	0.021
	Legitimate files	0.042(p)	0.008
Entropy test	Legitimate files	6.461	0.225
	Mimicry files	6.596	0.035
	Encrypted .pdf	7.934	0.043

Take the K-S test between mimicry executables and the raw binaries, also between mimicry executables and legitimate files. If the results of the test is low (lower than 0.05), then we can say that the two data set are semblable, otherwise, they are different. The mean score of p in formal test is 0.041 and the standard deviation is 0.021, while the score of p in later test is 0.092 and the standard deviation is 0.008. Therefore, the K-S test is unable to reliably differentiate mimicry files from legitimate files, while it can distinguish the mimicry files from the raw binaries easily, as shown in Figure 7 (a).

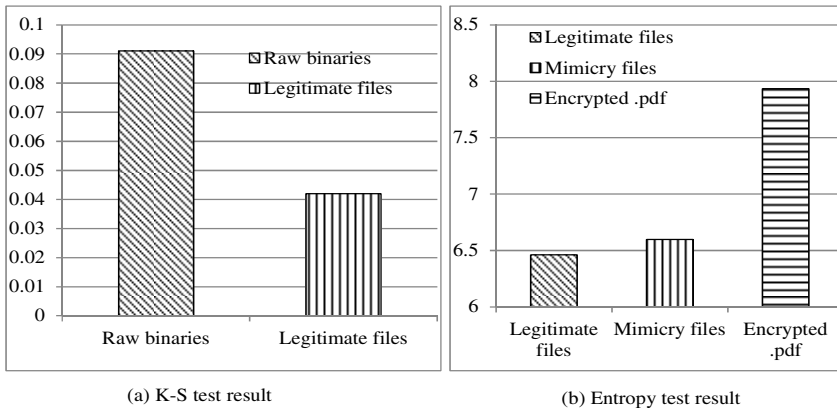


Fig. 7. Statistical test results in diagrammatical form

The mean and standard deviation of the byte entropy test scores are listed in Table 2. For the legitimate files, the mean entropy value is 6.461 and the standard deviation is 0.225. For mimicry files, the mean entropy value is 6.596, and the value is slightly higher on average, but fall well within one standard deviation of the legitimate mean.

Based on the results, the byte entropy test is unable to differentiate mimicry files from legitimate files. Moreover, packed and encrypted executables have byte entropies over 7 [7], so refined-mimimorphism is successful in digesting their packed content as normal executables, as shown in Figure 7 (b).

5 Conclusion

This paper has described a new approach to hide the statistical and behavior information of executable binary programs. It can transform a binary executable into a mimicry executable, with statistical and semantic characteristic highly similar to those of the mimicry target. Exploiting the method, special programs can successfully evade certain statistical anomaly detections.

We have implemented a prototype of the method on the X86 platform and windows 7 system. The experimental results validate its efficiency in evading statistical anomaly detection-the byte frequency distribution test and entropy test. In the test process, the mimicry output produced by the prototype falls within the ranges of normal executable files. Compared to the method of [5], refined-mimimorphism can effectively hide the statistical and semantic feature of binary program, which is as well as [5]. The strong-point of refined-mimimorphism is the low encoding error rate with little time sacrifice. In the future work, we are interested in the study of effective-measures against mimimorphism. For instance, we will explore the possibility of detecting mimimorphism through higher-order mimic functions and higher level semantics.

Acknowledgments. This research was supported partly by national science plan 2009AA01Z434 and university future fund F1201. Any opinions, findings, conclusions, or recommendations expressed in paper are of the authors except those cited in References.

References

1. Caballero, J., Grier, C., Kreibich, C., Paxson, V.: Measuring pay-per-install: The commoditization of malware distribution. In: Proceedings of USENIX Security Sym. (2011)
2. Szor, P.: The art of computer virus research and defense, pp. 67–72. Symantec Press (2005)
3. Lyda, R., Hamrock, J.: Using entropy analysis to find encrypted and packed malware. *IEEE Security and Privacy* 5(2), 40–45 (2007)
4. Vishwath, M.: Frankenstein: Stithching malware from benign binaries. In: Proceedings of the 6th USENIX Conference on Offensive Technologies (WOOT 2012), pp. 8–15 (2012)
5. Wu, Z., Steven, G., Xie, M.: Mimimorphism: A new approach to binary code obfuscation. In: Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS 2010), pp. 536–546 (2010), doi:10.1145/1866307.1866368
6. Wayner, P.: Mimic function. *Cryptoglia* 16(3), 193–214 (1992)
7. Lyda, R., Hamrock, J.: Using entropy analysis to find encrypted and packed malware. *IEEE Security and Privacy* 5(2), 40–45 (2007)

Glaucus: Predicting Computing-Intensive Program's Performance for Cloud Customers

Xia Liu^{1,*}, Zhigang Zhou¹, Xiaojiang Du², Hongli Zhang¹, and Junchao Wu¹

¹ School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, 150001, China
baleitu315@gmail.com,
{zhouzhigang, zhanghongli, wujunchao2011}@pact518.hit.edu.cn

² Department of Computer and Information Sciences,
Temple University, Philadelphia, PA 19122, USA
dux@temple.edu

Abstract. As Cloud computing has gained much popularity recently, many organizations consider transmitting their large-scale computing-intensive programs to cloud. However, cloud service market is still in its infant stage. Many companies offer a variety of cloud computing services with different pricing schemes, while customers have the demand of "spending the least, gaining the most". It makes a challenge which cloud service provider is more suitable for their programs and how much computing resource should be purchased. To address this issue, in this paper, we present a performance prediction scheme for computing-intensive program on cloud. The basic idea is to map program into an abstract tree, and create a miniature version program, and insert checkpoints in head and tail for each computable independent unit, which record the beginning & end timestamp. Then we use the method of dynamic analysis, run the miniature version program on small data locally, and predict the whole program's cost on cloud. We find several features which have close relationship with program's performance, and through analyzing these features we can predict program's cost on the cloud. Our real-network experiments show that the scheme can achieve high prediction accuracy with low overhead.

Keywords: Cloud computing, predict, cost, performance.

1 Introduction

Cloud computing has recently become an absorbing platform that attracted a substantial amount of attentions from both industry and academia [1-5]. It is a promising computing paradigm which provides a great platform with its spectacular storage capacity and powerful computational capability. As the advantage list of cloud, customers have motivated to outsource their computing-intensive applications to cloud. However, customers have no idea about how to buy resources of the cloud. So it has become a crucial challenge for the development of the cloud. Due to interest conflict, customers do not believe the resource purchase strategy given by cloud, while customers lack of professional knowledge. And one of the things they care about most is

how to run more applications spending less money meanwhile observing SLA. To our best knowledge, the existing methods can't resolve this crucial problem.

Realizing the problem above, we aim to develop a middleware to predict the resource cost of the customers' application on cloud without migrating the application. Taking use of the middleware, customers can know the general status of running an application, and this will guide customers to buy the cloud.

To build such a middleware, we first need to analyze customer's applications and create the miniature version of the application. Specifically, we transform the application into abstract tree and then we separate it into many computable independent units and compare the similarity among them. Based on this, we create a miniature version program. We use the method of dynamic analysis. By running the miniature version program on small dataset locally, we predict the performance of application on large dataset on the cloud. And we find several features that have close relationship with the performance of application. The experimental results show that our middleware can predict the performance accurately. Our contributions are summarized as follows.

- Our middleware can predict customers' applications performance, and provide reference for customers to run more applications spending less money.
- We predict the performance of applications without migrating any application to the cloud.

2 Framework

As shown in Fig. 1, there are three kinds of entities in this system: cloud, customers, and middleware. We define our system formally first. To predict applications' performance, middleware need three parts of input from customers and the cloud respectively. The data needed to be transported to the middleware is defined as a triple $T(D_A, D_T, f(D_T))$, where D_A refers to customer's application, D_T refers to the cloud type which customer wants to buy, and $f(D_T)$ refers to environmental difference between the cloud and local. After customer input cloud type D_T , middleware gets a five-tuple parameter $F(E_C, E_M, E_S, E_B, E_O)$ of the cloud, where E_C refers to CPU, E_M refers to memory size, E_S refers to storage space, E_B refers to bandwidth and E_O refers to operating system. Based on the information above, middleware can calculate environmental difference $f(D_T)$. Based on this, our middleware predict the performance of applications and return the result to customers to give them suggestion.

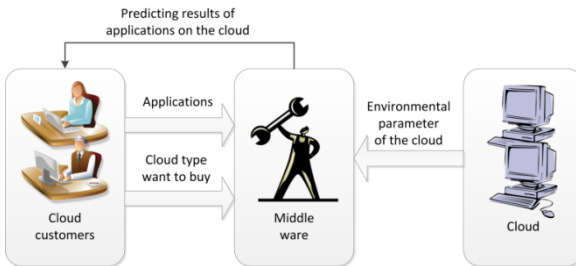


Fig. 1. Framework of middleware

3 Miniature Version Program

3.1 Abstract Tree

As our goal is to estimate customers' applications, we first need to analyze application. It is important to normalize application for the aim to analyze it easily, so we transform the application to abstract tree.

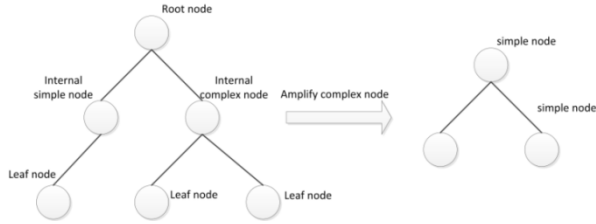


Fig. 2. Abstract tree

Table 1. Building abstract tree

Algorithm 1 Building Abstract Tree
INPUT: Customer's Application: tokens
OUTPUT: Abstrat Tree
MAIN
1: $n_{root} \leftarrow NULL, i = 0$
2: while token \neq NULL do
3: if token = "while" then
4: new $n_i = NULL, n_i = N_c$
5: $n_i.fathernode = n_{i-1}, i++$
6: else if token = "for" then
7: ...
8: else if $n_{i-1} = N_c$ & n_{i-1} is completed
&token \in simple token then
9: new $n_i = NULL, n_i = N_s$
10: $n_i.fathernode = n_{i-1}, i++$
11: else if token \in simple token then
12: merge(token, n_{i-1})
13: ...
14: end if
15: end while
16: return C

Here we give the definition of abstract tree. As Fig. 2 shows, there are three kinds of nodes in an abstract tree. Among which, root node N_R refers to the beginning of application, internal node N_I refers to different middle parts of application and leaf node N_L refers to end of application. Specifically, internal node includes two kinds of nodes which are called simple node N_S and complex node N_C . As we define, simple node refers to uncomplicated part of application (e.g., assignment statement), and complex node refers to complicated blocks in the application (e.g., loop, function call). At the same time, every complicated block also includes many simple nodes. That means we take every complicated block as a whole part “complex node”, meanwhile, the complicated block is also be analyzed and transformed to a sub-tree.

3.2 Independent Computable Unit

Based on abstract tree, we transform the whole program into a uniform structure. As our goal is to get miniature version program, we first split the main branch of abstract tree into independent computable units.

Definition 1. Independent Computable Unit: $U = \{u_i | u_i \in U, u_i \cap u_j = \emptyset\}$ Specifically, independent computable unit is a block of program which contains a series of successive statements, and there is no statement having relationship with this block outside it.

We find independent computable unit based on the abstract tree. First, we scan the abstract tree from bottom to top. If statements included in father node have semantic relation with the node which is being scanned, we merge father node with this node. Otherwise, father node belongs to a new independent computable unit.

3.3 Miniature Version Program

Although we get abstract tree, this branch is still too large to be analyzed, so we aim to get a miniature version of application.

Table 2. Typical Structures’ Running Time Analysis

Structure	Running time
While	$looptimes * \sum_{s_i \in \text{while} \& \text{judge}} get_time(s_i) + get_time(s_{judge})$
Switch	$get_time(s_{judge}) + \sum_{s_i \in \text{switch}} get_time(s_i)$
For	$looptimes * \sum_{s_i \in \text{For} \& \text{judge}} get_time(s_i) + get_time(s_{judge})$

We first separate main branch into several independent computable units, then reduce each unit. As different unit has different structure and also has different importance, if we want to get a more accurate miniature version program, we should not reduce each unit according to the unified standard. Therefore, we use the method of

analyzing running time ratio of different program structure, and then define the reducing ratio of each unit. As Table 2 illustrates, we analyze reducing ratio taking three typical structures.

Based on the definition of different structures’ running time formula, we will figure out the whole structure’s running time if we know each statement’s running time. Then we split every statement into machine language. As Table 3 shows, we get the CPU cycles of machine instruct.

Table 3. The CPU Cycles Of Machine Instruct

Index	CPU cycles	Instruct
23/0	= 2	;AND r,r
23/1	= 6	;AND r,m
31/1	= 7	;XOR m,r
33/0	= 2	;XOR r,r
42	= 2	;INC eDX
28/0	= 2	;SUB r,r8
7C	= 7	;JL rel8

Based on the analysis above, we know CPU cycles that each statement need. So we define unit’s reducing ratio as the inverse ratio of CPU cycles of each unit. Then we reduce each unit according to reducing ratio. Specifically, we reduce each unit’s number of statements in terms of reducing ratio of the unit, then we get miniature version program. Formula (1) expresses the cost of the miniature version.

$$C_m = C_r * \varphi * \theta \tag{1}$$

So we enlarge cost results of program’s cost as follows:

$$C_r = (C_m / \theta) * \varphi^{-1} \tag{2}$$

where C_r is real cost of program, C_m is miniature version program’s cost, φ is complexity function of different structure and θ is reducing ratio of each unit.

4 Predicting Program’s Performance

4.1 Core Variable Set

After getting the miniature version program, it is crucial to analyze it to find out the relationship between application and resources it have to occupy on the cloud.

As the complex node takes most of running time and also more other resources, we mainly analyze the complex node. Through analyzing complex node, we find out that in the complex node, what have close relationship with the resources the application occupies are some statements that contain the special variables. We define these variables as core variables, which are the variables which have high weight because they

appear in the miniature version program more frequently or appear in some important statements (e.g. malloc statement).

Definition 2. Core variable set: $V = \{v_i \mid v_i \in V, w_{v_i}/w_{tree} > \alpha, i \in [1, n] \text{ and } 0 < n \leq \max_num(N_v)\}$, where v_i is one of the core variables which are in the miniature version program, w_{v_i} is weight of v_i , w_{tree} is total weight of miniature version program tree, α is the threshold which limit the standard weight of v_i , n is number of core variables and $\max_num(N_v)$ is the total number of the variables appearing in miniature version program.

To find out core variables, we need to calculate weight of variables in the complex node. As described above, we have already known the CPU cycles each statement need, so we define statement's weight as corresponding CPU cycles. Then we calculate variable's weight according to the weight of statements which contain the variable. The weight of variable v_i is given by:

$$w_{v_i} = \sum w_{s_i} * n_{s_i} \quad (3)$$

where w_{s_i} is weight of statements which contain variable v_i and n_{s_i} is number of times these statements appears in the miniature version program.

After getting every variable's weight, we sort these variables and find out all of the variables whose weight are bigger than α . According to this method, we get m core variables.

4.2 Dynamic Analysis

As the performance of an application has great relationship with frequency of memory page swapping, we analyze several elements influencing memory swapping. We analyze two important features that are close related to the core variables, which are variables' frequency V_F , and average density V_D . Specifically, the frequency of core variables refers to the percentage of times a core variable occurring in the miniature version program, and average density refers to average distance of statements which contain a specific core variable.

Through the method of dynamic analysis, we find that when given fixed environmental parameters, performance of program will reach bottleneck as dataset becomes larger. So we analyze performance of program from two aspects. Before program's performance reaches bottleneck, we find that cost of program increases linearly approximately. We first run miniature version program on a specific small dataset, and get two pairs of program's cost (i.e. time cost and memory cost). Then predict cost of program before program's performance reaching bottleneck. The memory cost of program is given by:

$$m_s = \frac{M_2 - M_1}{D_2 - D_1} * d \quad (4)$$

where m_s refers to memory resource when running program on the specific small dataset, d is scale of dataset, M_1 and M_2 are memory size of program when running the program on two specific data scale respectively, and D_1 and D_2 are corresponding data scale.

As we know the maximal memory size M_{max} of our platform, we can figure out corresponding data scale when the program taking use of all the memory. So the largest data scale before program's performance reaching bottleneck is given by:

$$D_{max} = \frac{M_{max} * (D_2 - D_1)}{M_2 - M_1} \quad (5)$$

And by the same method, we can figure out running time of program:

$$t = \frac{T_2 - T_1}{D_2 - D_1} * d \quad (6)$$

where t_s refers to running time when running program on the specific small dataset, and T_1 and T_2 are running time of program when running the program on two specific data scale respectively.

Based on analysis above, we figure out the running time T_{max} of program when program's performance reaching bottleneck, which is given by:

$$T_{max} = \frac{T_2 - T_1}{D_2 - D_1} * \frac{M_{max} * (D_2 - D_1)}{M_2 - M_1} \quad (7)$$

Then we know the inflection point which represents the bottleneck of program's performance when given a fixed environmental parameter. After that, we predict performance of program based on the two important features. As we know, different features have different importance, we define the two features weight as w_{V_F} , and w_{V_D} respectively. And the running time of program on big dataset is given by:

$$t_b = \frac{T_2 - T_1}{D_2 - D_1} * d + (w_{V_F} V_F + w_{V_D} V_D) T_{max} \quad (8)$$

And memory resource that program occupy is given by:

$$m_b = \frac{M_2 - M_1}{D_2 - D_1} * d + (w_{V_F} V_F + w_{V_D} V_D) M_{max} \quad (9)$$

Based on our analysis of program's cost locally, we can figure out the whole program's performance on the cloud. We use the following performance model as our final performance model:

$$\Omega = \Upsilon \lambda \mu \quad (10)$$

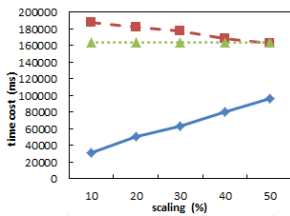
where Ω refers to the predicting value of program's cost on the cloud, Γ is the cost of miniature version program on our local middleware platform, λ is environmental parameter difference degree between local and cloud, and μ is the reducing ratio between real program and the miniature version program.

5 Evaluation

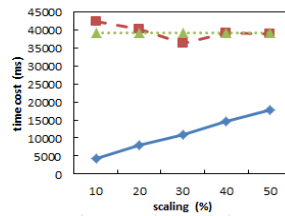
To evaluate the performance of our scheme, we run real computation-intensive application on both local platform and the real cloud platform. Based on this, we compare the results of application's cost of the two platforms. The configuration of our experiments is as follows:

Our local platform is a PC with Ubuntu 12.04 CPU, 512 MB Memory and 20GB hard disk. And we build our cloud test platform with 8-core 2.93 GHz Intel Xeon CPU, 24 GB memory, 500 GB and installed with Linux 2.6.18.

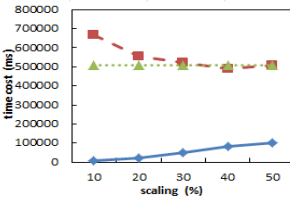
To find the reducing scale of miniature version program, we do experiment based on four typical computing-intensive programs (i.e. matrix multiplication, red-black tree, heap sort, maximal sum of submatrix). As Fig. 3 shows, when changing scale of program, the predicting of original program's performance also changes. And when reducing scale percentage is 20%, prediction result is the best.



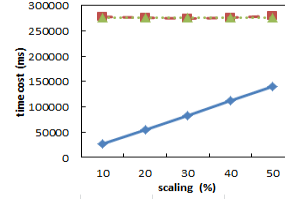
(a) matrix multiplication



(b) red-black tree



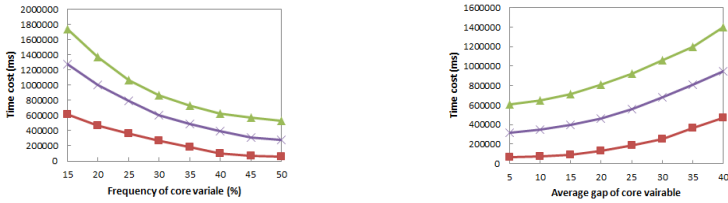
(c) heap sort



(d) maximal sum of submatrix

Fig. 3. Changing reducing scale of miniature version program

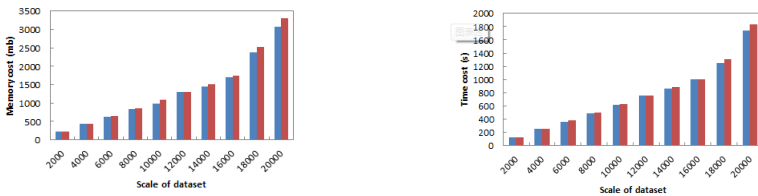
To prove the two features(i.e. frequency and average gap) have great influence on program's performance, we run our benchmark program locally, and change frequency and average gap of a specific core variable under the condition that the program filling all the available memory. And we find out that program's running time changes obviously with the variety of the two features. As Fig. 4(a) shows, program's time cost changes sharply as frequency of core variable changes from 15% to 50%. And Fig. 4(b) illustrates when changing average gap of a fixed core variable from 5 to 40, program's time cost also changes greatly.



(a) Changing frequency of core variable (b) Changing average gap of core variable

Fig. 4. Application’s cost changes with features

Finally, we run our benchmark program on our local middleware platform with the dataset scale of 2000 and 4000 respectively. Then we predict cost of program on large dataset scale up to 20000 based on our scheme, and also run the program on cloud test platform. As Fig. 5 shows, our middleware can accurately predict the performance of computing-intensive applications on cloud with small dataset scale.



(a) Prediction result of program’s memory cost (b) Prediction result of program’s time cost

Fig. 5. Prediction result of program’s cost

6 Related Work

Many efforts have been made to predict performance of the computing-intensive program. Li et al. [6, 7] compares performance of multiple cloud servers. This work is used to predict the performance of CPU-intensive applications across a large collection of CPU types and gives purchase suggestions for computing-intensive tasks and Web application. Zhang et al. [8] presents a performance prediction scheme that constructs a miniature version program to run in local machines, and then replays it in cloud to get the performance ratio between local and cloud. Ye et al. [9] proposes a testing algorithm based on HyperSentry [10] to detect SLA violations of physical memory size in virtual machine (VM). Sommers et al. [11] proposes a novel active measurement methodology to monitor whether the characteristics of measured network path are in compliance with performance targets specified in SLAs. Study [12] proposes a new passive traffic analysis method for on-line SLAs assessment, which reduces both the need for measuring QoS metrics as well as the interactions between the ingress and egress nodes in the network.

7 Conclusion

In this paper, we propose a method of predicting computing-intensive program's performance on cloud. We identified and addressed three key challenges: (1) how to find unified form of computing-intensive applications for analyzing it easier, and (2) how to predict the performance of applications on large scale dataset without migrating any application to the cloud (3) how to provide reference for customers to run more applications spending less money. Our experiments showed that our scheme can achieve accurate prediction results with low overhead.

Acknowledgment. This work is supported by National Basic Research Program (973 Program) of China (2011CB302605), the project of National Natural Science Foundation of China (60903166, 61100188, 61173144), and National High Technology Research and Development Program (863 Program) of China (2011AA010705).

References

1. Amazon Web Service, <http://aws.amazon.com/>
2. Google AppEngine, <http://code.google.com/appengine/>
3. Armbrust, M., Fox, R.G.A., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A.: Stoica, I., Zabarria, M.: Above the Clouds: A Berkeley View of Cloud Computing. University of California, Berkeley, Tech. Rep. (2009)
4. Sripanidkulchai, K., Sahu, S., Ruan, Y., Shaikh, A., Dorai, C.: Are Clouds Ready for Large Distributed Applications? In: Proc. SOSP LADIS Workshop (2009)
5. Microsoft Windows Azure, <http://www.microsoft.com/>
6. Li, A., Liu, X., Yang, X.W.: CloudCmp: Comparing Public Cloud Providers. In: USENIX/ACM Symposium on Networked Systems Design and Implementation (April 2011)
7. Li, A., Liu, X.: CloudCmp: Shopping for a Cloud Made Easy. In: 2nd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud (2010)
8. Zhang, H.L., Li, P.P., Zhou, Z.G., Du, X.J., Zhang, W.Z.: A Performance Prediction Scheme for Computation-Intensive Applications on Cloud. In: Proc. of ICC 2013 (2013)
9. Ye, L., Zhang, H.L., Shi, J.T., Du, X.J.: Verifying Cloud Service Level Agreement. In: Proc. of Globecom 2012 (2012)
10. Azab, A.M., Ning, P., Wang, Z., Jiang, X., Zhang, X., Skalsky, N.C.: HyperSentry: Enabling Stealthy In-Context Measurement of Hypervisor Integrity. In: Proc. of the CCS 2010, Chicago, Illinois, pp. 38–49 (2010)
11. Sommers, J., Barford, P., Duffield, N., Ron, A.: Multi-objective Monitoring for SLA Compliance. IEEE/ACM Transactions on Networking 18(2), 652–665 (2010)
12. Serral-Gracia, R., Yannuzzi, M., Labit, Y., Owezarski, P., Masip-Bruin, X.: An Efficient And Lightweight Method for Service Level Agreement Assessment. Computer Networks 54(17), 3144–3158 (2010)

A Multiwavelet Support Vector Machine Prediction Algorithm for Avionics PHM*

Xinzhou Zhou¹, Zheng Xiang², Meng Liu², and Jiang Xiang²

¹ Science and Technology on Avionics Integration Laboratory, Shanghai, China
zhouxzz@163.com

² School of Telecommunication, Xidian University, Xi'an, China
zhx@mail.xidian.edu.cn,
aileen127@163.com,
xiangjianghappiest@hotmail.com

Abstract. To improve the accuracy of the prediction in avionics prognostics and health management (PHM), a variety of theories and methods are studied. In this paper, a prediction algorithm based on multiwavelet support vector machine (WSVM) is proposed. Multiwavelet denoising is used for signal data preprocessing. Then multiwavelet is employed to decompose the data into several subsequences at different scales. These subsequences are predicted by different support vector machines respectively. Finally, the final predicted results reconstituted from the subsequences are obtained. To validate the model, experiment data from a set of certain avionics voltage data is used. Predicted results of the proposed algorithm are validated to be more accurate than that of traditional support vector machine prediction algorithm. The mean square error (MSE) is decreased to 0.1956.

Keywords: Prognostics and health management (PHM), Multiwavelet transform, Support vector machine (SVM), Prediction.

1 Introduction

Since Prognostics and Health Management (PHM) [1-2] technology is becoming an important part of the new generation of aircraft in design and use, building avionics PHM system is a new direction for development of the aircraft fault diagnosis system. Different PHM systems are developed for different types of system applications in the national defense military, aerospace and other industrial areas; however, the same basic idea of the design is utilized. The basic structure of the avionics PHM system is shown in Fig.1 [3].

As a core support technique of PHM, prognostic techniques have been widely researched by academia and industry. For prognostic techniques, some researchers have focused on signal prediction based on auto regression (AR) and moving average

* This work was supported by Science and Technology on Avionics Integration Laboratory and Aeronautical Science Fundation of China under Grant 20105581016.

(MA) [4], physical model, Kalman filter[5], extended Kalman filter, particle filter, and expertise-based methods. All these methods are classified as model-based prognostic techniques [6] that can realize real-time prediction. However, due to actual project accuracy, the practical application of model-based prognostic techniques is limited at the scope and effect.

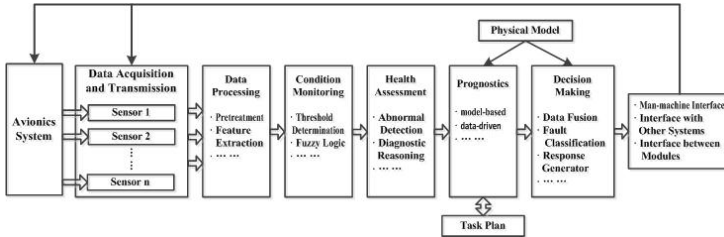


Fig. 1. Basic structure of the avionics PHM system

Neural network is the most widely used method in prognostic and applied research. Different from the model-based techniques, neural network is a data-driven and self-adaptive analysis approach. Along with the progress of the research, many improved algorithms of neural network are proposed, such as wavelet neural network [7] (WRY), fuzzy neural network [8] (FNN), etc. Support vector machine (SVM) and neural network are both data-driven prognostic methods. In contrast to neural network [9-11], SVM accounts for the trade-off between learning and generalizing abilities by minimizing the structural risk [12]. The SVM learning approach is popular because it exhibits desirable properties, such as good generalization performance and over fitting prevention capabilities [13-15]. Thus, in recent years, SVM has been proven to be an effective and generalized method for prognostic.

The purpose of this paper is to introduce a PHM prediction algorithm integrating multiwavelet and the least squares support vector machine (LS-SVM). Different from wavelet kernel SVM [16], the new approach takes advantage of the good frequency resolution of multiwavelet, and predicts on each scale by using SVM. In this algorithm, multiwavelet is not only employed to remove the noise, but to decompose the data into several scales, preparing for LS-SVM to predict respectively. The block diagram of the algorithm is shown in Fig.2.

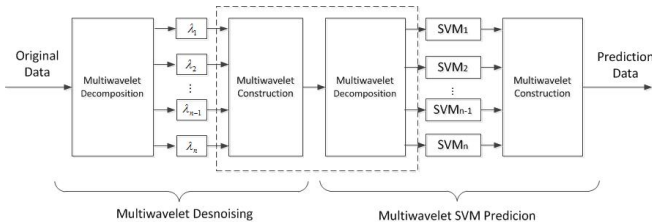


Fig. 2. Block diagram of algorithm

The paper is organized as follows. In Section 2, the multiwavelet threshold denoising method is used to deal with the noise data. Then multiwavelet combined with LS-SVM accomplishes the prediction. Section 3 illustrates the feasibility and effectiveness of the algorithm with simulation and analysis. Finally, Section 4 concludes with a summary.

2 The Proposed Algorithm

2.1 Multiwavelet Denoising

During transmission and reception, signals do not exist without noise. Therefore an effective means of removing the noise would be an invaluable first stage for many further data analysis.

Wavelet thresholding methods [17] for denoising, in which the wavelet coefficients are processed in order to remove their noisy part, were first introduced by Donoho. These denoising methods do not require any particular assumptions about the nature of the original signal, permits discontinuities and spatial variation in the signal, and exploits the spatially adaptive multiresolution of the wavelet transform. In this paper, a multiwavelet thresholding method is adopted to remove the noise. In multiwavelet denoising method, the thresholds of each layer can be set as different values according to the actual situation. Therefore, denoising based on multiwavelet transform not only has the advantages of traditional wavelet denoising methods in integrality, but has the characteristics of higher flexibility and practicability.

The observed signals are as follows:

$$f(t) = s(t) + n(t) \tag{1}$$

Where $s(t)$ is the original signal, $n(t)$ is the noise.

An N-point discrete signal $f_n, n = 0, 1, \dots, N - 1$ is obtained by sampling $f(t)$. The objects of the multiwavelet decomposition and reconstruction are vector sequences, so the original data should be converted into a vector sequence by the pre-filer Q before calculation, as shown in Fig.3. Filter matrixes H and G have a variety of choice. In this algorithm, GHM multiwavelet which has short support, (anti)symmetry, orthogonality and vanishing moment with order 2 is adopted.

The decomposition and reconstruction functions are,

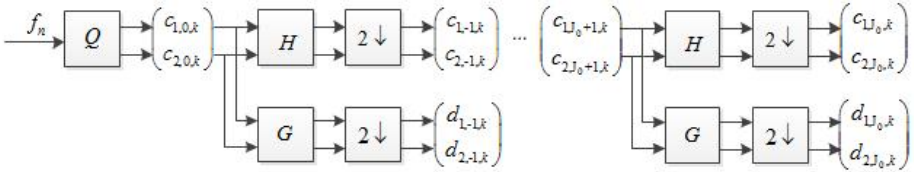
$$c_{j-1,k} = \sqrt{2} \sum_n H_n c_{j,2k+n} \tag{2}$$

$$d_{j-1,k} = \sqrt{2} \sum_n G_n c_{j,2k+n} \tag{3}$$

$$c_{j,n} = \sqrt{2} \sum_k (H_k^* c_{j-1,2k+n} + G_k^* d_{j-1,2k+n}) \tag{4}$$

Where H, G, H^* and G^* are the filter matrixes of the multiwavelet transformation, $c_{j,k} = (c_{1,j,k}, \dots, c_{r,j,k})^T, d_{j,k} = (d_{1,j,k}, \dots, d_{r,j,k})^T, r = 1, 2, j = 0, 1, 2, 3$.

Multiwavelet Decomposition



Multiwavelet Reconstruction

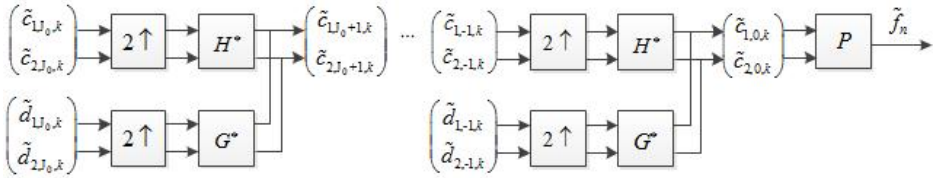


Fig. 3. multiwavelet decomposition and reconstruction

In order to simplify the discussion, multiwavelet coefficients $c_{i,j,k}$ and $d_{i,j,k}$ are collectively referred to as $w_{i,j,k}$.

After several times of the multiwavelet decomposition, some multiwavelet coefficients have relatively large values in certain positions, which correspond to the important information of the original signal, while the most of the other positions have small values. The multiwavelet coefficients of the white noise are uniformly distributed in each scale, and the coefficient values are reduced with increasing scales. Therefore, the usual way of denoising is to find a suitable value as the threshold λ . The coefficients (mainly caused by noise) lower than λ are set to zero, while the multiwavelet coefficients higher than λ (mainly caused by the signal) are retained. In this paper, a compromised algorithm between soft and hard threshold is employed to remove the noise. The threshold λ_l can be calculated with Eq.5,

$$\tilde{w}_{i,j,k} = \begin{cases} \text{sign}(w_{i,j,k}) \cdot (|w_{i,j,k}| - \alpha \lambda_l), & |w_{i,j,k}| \geq \lambda_l, \\ 0, & |w_{i,j,k}| < \lambda_l, \end{cases} \quad 0 \leq \alpha \leq 1 \tag{5}$$

The parameter α can be adjusted between 0 and 1 to achieve a better denoising performance.

Finally, all the processed $\tilde{w}_{i,j,k}$ at each scale reconstruct the vector sequence.

Then denoising result \tilde{f}_n is restored by the post-filter P .

2.2 Multiwavelet Support Vector Machine

In prediction part, first of all, the data \tilde{f}_n is decomposed, then SVM on each frequency band predicts respectively, and the resulting data reconstruct at last. As shown in Fig.2, when the selected decomposition scales are the same, the reconstruction for denoising and the decomposition for prediction is a reciprocal process. So the model is simplified as shown in Fig.4. First, decompose the original data f_n and obtain the multiwavelet coefficients $w_{i,j,k}$; then, process the coefficients by using different thresholds λ_j ; next, use SVM to predict the coefficients $\tilde{w}_{i,j,k}$ at each scale; lastly, reconstruct the prediction data \hat{f}_{n+l} .

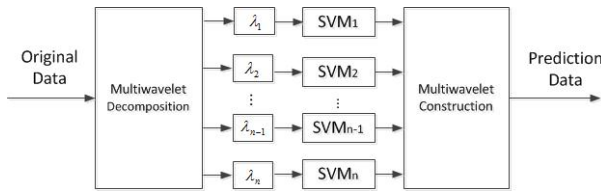


Fig. 4. Simplified prediction model

LS-SVM is employed to predict the multiwavelet coefficients at each scale. Like the SVM, the LS-SVM is based on the margin-maximization principle that performs structural risk minimization; moreover, it inherits the SVM generalization capacity. In LS-SVM, there are two important parameters, one is the regularization parameter, which depends on the degree of trade-off between the training error minimization and the smoothness; and the other parameter σ^2 , usually in the Gaussian RBF kernel, is squared bandwidth. The cross-validation method is used to determine the optimal values of these two parameters in training set.

The basic idea of LS-SVM is a nonlinear map defined by dot products function. Data are mapped into some high dimensional feature space.

The multiwavelet coefficients can be represented as $\{\tilde{w}_{i,j,1}, \tilde{w}_{i,j,2}, \dots, \tilde{w}_{i,j,N}\}$, and the matrix for LS-SVM model can be constructed as Eq.6 and Eq.7, establishing m -order prediction model.

$$X = \begin{bmatrix} \hat{w}_{i,j,1} \\ \hat{w}_{i,j,2} \\ \vdots \\ \hat{w}_{i,j,N-m} \end{bmatrix} = \begin{bmatrix} \tilde{w}_{i,j,1} & \tilde{w}_{i,j,2} & \cdots & \tilde{w}_{i,j,m} \\ \tilde{w}_{i,j,2} & \tilde{w}_{i,j,3} & \cdots & \tilde{w}_{i,j,m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{w}_{i,j,N-m} & \tilde{w}_{i,j,N-m+1} & \cdots & \tilde{w}_{i,j,N-1} \end{bmatrix} \tag{6}$$

$$Y = \begin{bmatrix} \tilde{w}_{i,j,m+1} \\ \tilde{w}_{i,j,m+2} \\ \vdots \\ \tilde{w}_{i,j,N} \end{bmatrix} \tag{7}$$

Where m denotes the embedded dimension, X is the training set, and Y is the prediction set.

By training, the obtained LS-SVM model can be written as,

$$y_{i,j,n} = \sum_{s=1}^{N-m} (\bar{\alpha}_{i,j,s}^* - \bar{\alpha}_{i,j,s}) K(\bar{w}_{i,j,s}, \bar{w}_{i,j,n}) + b_{i,j} \tag{8}$$

$$K(\bar{w}_{i,j,s}, \bar{w}_{i,j,n}) = \exp\left\{-\frac{|\bar{w}_{i,j,s} - \bar{w}_{i,j,n}|^2}{\sigma^2}\right\} \tag{9}$$

Where $n = m + 1, \dots, N$, $K(\bar{w}_{i,j,s}, \bar{w}_{i,j,n})$ is the Gaussian RBF kernel function, $\bar{\alpha}_{i,j,s}^*$, $\bar{\alpha}_{i,j,s}$, $b_{i,j}$ are parameters of the model obtained by training.

Since $\hat{w}_{i,j,t-m+l} = \{\tilde{w}_{i,j,t-m+1}, \tilde{w}_{i,j,t-m+2}, \dots, \tilde{w}_{i,j,t+l-1}\}$, l -th step forecast is:

$$\begin{aligned} \hat{w}_{i,j,n+l} &= \sum_{s=1}^{N-m} (\bar{\alpha}_{i,j,s}^* - \bar{\alpha}_{i,j,s}) K(\bar{w}_{i,j,s}, \bar{w}_{i,j,n-m+l}) + b_{i,j} \\ &= y_n \end{aligned} \tag{10}$$

After the respective threshold processing at different scales, $\tilde{w}_{i,j,k}$ is used as training data and the input of SVM to perform predictive processing, resulting in $\hat{w}_{i,j,n+l}$.

Finally, the final predicted value \hat{f}_{n+l} is reconstructed.

3 Algorithm Flow and Simulation

3.1 Algorithm Flow

According to the presentation and analysis of the proposed algorithm, the specific simulation algorithm flow chart is shown in Fig.5.

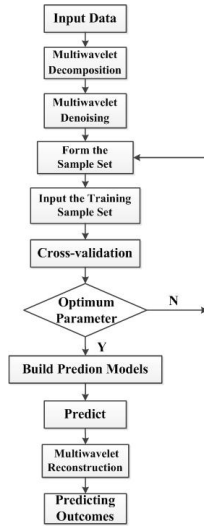


Fig. 5. Flow chart of prediction mode

The multiwavelet support vector machine prediction algorithm process is as follows:

Step 1: Initiate $n = 1$, and use GHM multiwavelet to decompose the original data set $\{f_n, \dots, f_{n+L}\}$ into several decomposition coefficients $\{w_{i,j,n}, \dots, w_{i,j,n+L}\}$, where L is the training data length. Then carry out the threshold processing to get $\{\tilde{w}_{i,j,n}, \dots, \tilde{w}_{i,j,n+L}\}$.

Step 2: Let $\{\tilde{w}_{i,j,n}, \dots, \tilde{w}_{i,j,n+L-1}\}$ be the training set and input $\tilde{w}_{i,j,n+L}$ to predict $\hat{w}_{i,j,n+L+1}$. Then reconstruct results and calculate the final predicted value \hat{f}_{n+L+1} .

Step 3: Add f_{n+L+1} into the original data set. Let $n = n + 1$, and repeat step 2, ending the prediction until $n = N - 1$, where N is the length of the data for prediction.

3.2 Simulation

The voltage of the force sensor on the uniform rotation test equipment is sampled, whose variation range is 5~5V. The voltage value in the ideal case should be a sine curve, but in fact, the actual signal is always accompanied with the noise and other interference.

For the accuracy of a prediction model, Mean Square Error (MSE) is often used to calculate the prediction error:

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (s_i - \hat{f}_i)^2} \tag{11}$$

Where s_i is the ideal original data, \hat{f} is the predicted value.

To select an appropriate decomposition scale is a key point in multiwavelet transform. After a large number of pre-calculation and experiment, it proved that the data decomposition to scale 4 is more appropriate. From Fig.6, it can be seen that the low frequency part reflects the original characteristics of the signal, while the high-frequency part reflects the random, non-regularity characteristic of noise.

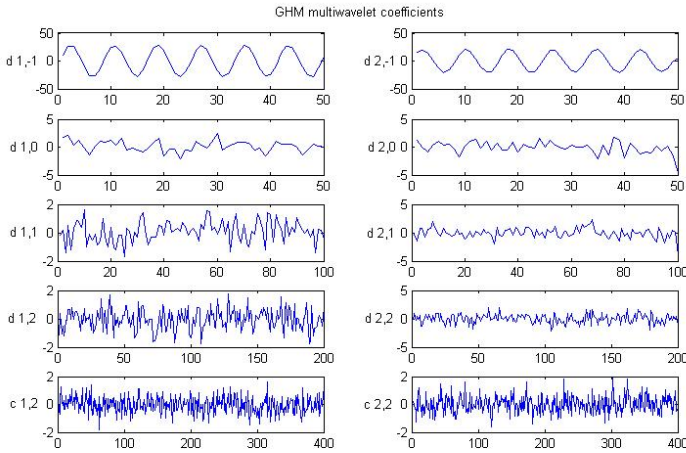


Fig. 6. GHM multiwavelet coefficients

As shown in Fig.7, compared with the data before denoising, the data processed by the multiwavelet threshold denoising retained most of the characteristics of the original data, and meanwhile the noise signal was effectively removed.

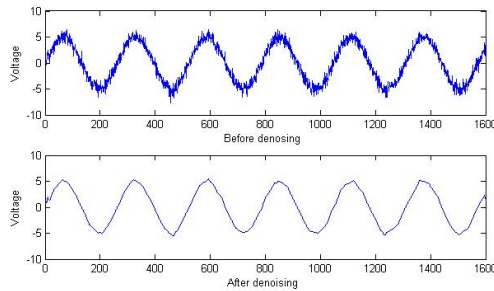


Fig. 7. Multiwavelet threshold denoising

To verify that the validity and feasibility of the proposed algorithm, predicted results of the median filter and LS-SVM algorithm and results of the proposed algorithm are compared here. For one step prediction, the predicted results are shown in Fig.8.

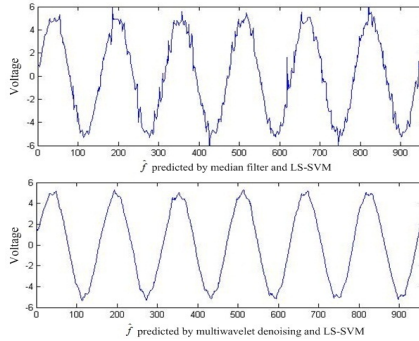


Fig. 8. Comparison of two algorithms

Seeing from Fig.8, the curve of the proposed algorithm is smoother and more close to the original signal $s(t)$. The mean square error of MSVM algorithm is 0.1956, while it is 0.4217 when using the median filter and LS-SVM algorithm. Therefore, the conclusion can be drawn that the proposed algorithm has high precision and good ability in predicting the force sensor voltage data. On the basis of the very good time frequency domain features of multiwavelet, the noise is removed effectively, and data features at each scale can be utilized by LS-SVM to obtain a better prediction effect. The excellent abilities of learning and generalization of LS-SVM is not ignorable. The decomposition scale and parameters of LS-SVM are related with the predicted results closely.

4 Conclusions

In this paper, a MSVM algorithm for prediction is proposed which shares good performance in denoising of multiwavelet, and excellent predictive ability of SVM. And the simulation and experiment of the proposed algorithm based on a set of certain voltage data proves that it is feasible and accuracy. There must have some future works to be done, such as exploring an effective way to design certain kernel functions for more complicated avionics equipment.

References

1. Hess, A., Fila, L.: The joint strike fighter (JSF) PHM Concept: Potential Impact on Aging Aircraft Problems. In: Proceedings of IEEE Aerospace Conference, vol. 6, pp. 3021–3026 (2002)

2. Keith, M.J., Raymond, R.B.: Diagnostics to Prognostics. A Product Availability Technology Evolution. In: The 53rd Annual Reliability and Maintainability Symposium(RAMS 2007), pp. 113–118 (2007)
3. Sun, B., Kang, R., Xie, J.: Research And Application of The Prognostic And Health Management System. *Systems Engineering and Electronics* 29(10), 1762–1767 (2007)
4. Mbamalu, G.A.N., El-Hawary, M.E.: Load Forecasting Via Suboptimal Seasonal Autoregressive Models And Iteratively Reweighted Least Squares Estimation. *IEEE Transactions on Power Systems* 8(1), 343–348 (1993)
5. Park, D.C., El-Sharkawi, M.A., Marks, R.J., Atlas, I.L.E., Damborg, M.J.: Electric Load Forecasting Using An Artificial Neural Network. *IEEE Transactions on Power Systems* 6(2), 442–449 (1991)
6. Michael, G.P.: Prognostics And Health Management of Electronics, pp. 3–20. John Wiley&Sons, Inc. (2008)
7. Yam, R., Tse, P., Li, L.: Intelligent Predictive Decision Support System for Condition-Based Maintenance. *International Journal of Advanced Manufacturing Technology* 17(5), 383–391 (2001)
8. Wang, P., Gachtsevanos, V.: Fault Prognostics Nsing Dynamic Wavelet Neural Networks. *AI EDAM Artificial Intelligence for Engineering Design Analysis and Manufacturing* 15, 349–365 (2001)
9. Aminian, M., Aminian, F.: A Modular Fault-Diagnostic System for Analog Electronic Circuits Using Neural Networks With Wavelet Transform as a Preprocessor. *IEEE Trans. Instru.* 56(5), 1546–1554 (2007)
10. Chen, C., Zhang, B., Vachtsevanos, G., Orchard, M.: Machine Condition Prediction Based on Adaptive Neuro-Fuzzy And High-Order Particle Filtering. *IEEE Transactions on Industrial Electronics* 58(9), 4353–4364 (2011)
11. Chen, C., Brown, D., Sconyers, C., Zhang, B., Orchard, M.E.: A NET Framework for An Integrated Fault Diagnosis And Failure Prognosis Architecture. In: *IEEE Autotestcon*, Orlando, FL, pp. 1–6 (2010)
12. Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: *Least Squares Support Vector Machines*. World Scientific, Singapore (2002)
13. Long, B., Tian, S.L., Wang, H.J.: Least Squares Support Vector Machine Based Analog Circuit Fault Diagnosis Using Wavelet Transform as Preprocessor. In: *ICCCAS 2008* (2008), doi:10.1109/ICCCAS.4657943
14. Cui, J., Wang, Y.R.: A Novel Approach of Analog Circuit Fault Diagnosis Using Support Vector Machines Classifier. *Measurement* 44, 281–289 (2011)
15. Long, B., Tian, S.L., Miao, Q., Pecht, M.: Research on features for diagnostics of filtered analog circuits based on LS-SVM. In: *IEEE Autotestcon*, Baltimore, MD, pp. 360–366 (2011)
16. Donoho, D.: Denoising by Soft-thresholding. *IEEE Trans on Information Theory* 41(3), 613–627 (1995)
17. Zhang, L., Zhou, W., Jiao, L.: Wavelet Support Vector Machine. *IEEE Trans. Systems, Man, and Cybernetics - Part B: Cybernetics* 34(1) (2004)
18. Goodman, T., Lee, S.: Wavelets of Multiplicity. *Trans. on Amer. Math. Soc.* 342, 307–324 (1994)

Using a Real-Time Top-k Algorithm to Mine the Most Frequent Items over Multiple Streams

Ling Wang^{1,*}, Zhao Yang Qu¹, Tie Hua Zhou², and Keun Ho Ryu²

¹Department of Computer Science and Technology, School of Information Engineering,
Northeast Dianli University, Jilin, China
smile2867ling@gmail.com, qzywww@mail.nedu.edu.cn

²Database/Bioinformatics Laboratory, School of Electrical & Computer Engineering,
Chungbuk National University, Chungbuk, Korea
{thzhou, khryu}@dblab.chungbuk.ac.kr

Abstract. Some applications such as sensor networks, internet traffic analysis, location-based services, and health measurements are always required for considering unbounded, fast, large-volumes, continuous, even for distributed stream data. It's a better way to use synopsis as a list of partial summaries of unknown item sets in order to reduce the memory space usage, let it can afford to process so fast and huge incoming data. Normally, different quantity of item set leads to different summaries, especially for Top-k operator which as a partial preprocess over synopsis. Therefore, we proposed smooth synopsis that dynamically assigns a numeral interval to resolve the items set, in order to maintain a more accurate approximate answers' list from partial Top-k processing. In particular, we proposed an algorithm (called SFI algorithm) to mine the most frequent items by a more adaptive and fast way in specific stream resources. Finally, our experimental results demonstrate the accuracy and efficiency of our approximation techniques.

Keywords: sliding window, synopsis, streaming data, Top-k query, SFI algorithm.

1 Introduction

There has been a concerted effort in recent years to build data stream management systems (DSMSs), either for general purpose or for specific streaming applications [1, 2, and 3]. The unique properties of data streams generate many interesting research problems in algorithm analysis and data management. By nature, a stored data set is appropriate when significant portions of the data are queried again and again, and updates are small or relatively infrequent. In contrast, a data stream is appropriate when the data is changing constantly (often exclusively through insertions of new elements), and it is either unnecessary or impractical to operate on large portions of the data multiple times. Therefore, in real data stream applications, data arrives very

* Corresponding author.

fast and the rate is so high that one may not wish to store all the data; yet, the need exists to query and analyze this data. One technique for producing an approximate answer to a data stream query is to evaluate the query not over the entire past history of the data streams, but rather only over sliding windows of recent data from the streams [4]. Most importantly, it emphasizes recent data, which in the majority of real-world applications is more important and relevant than old data: if one is trying in real-time to make sense of network traffic patterns, or phone call or transaction records, or scientific sensor data, then in general insights based on the recent past will be more informative and useful than insights based on stale data.

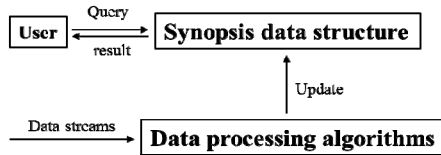


Fig. 1. Stream processing

Real-time analysis of network traffic has been one of the primary applications of data stream management systems. It concerns statistical analysis of data streams with a focus on newly arrived data and frequently appearing packet types. For instance, an ISP may be interested in monitoring streams of IP packets originating from its clients and identifying the users who consume the most bandwidth during a given time interval; see [5, 6] for additional motivating examples. It's more challenging for Top-k queries in stream data analysis because of only summarized results could be maintained in main memory. See figure 1, synopsis data structure can be used for managing approximate results after pre-processing by using some effective data processing algorithms such as aggregates. For example, suppose some packet streams represented as $\{a, b, c, d, e \dots\}$, then count all of them and response to Top-3 query of which are the most 3 appeared packets. Normally, a synopsis always manages some recent approximated results that how many times appeared for each packet type in a time or numeral interval. Actually, it's also a big job to identify frequency items over the whole data stored in memory. We proposed SFI algorithm to count items after creating some timestamp-bins (e.g., t_1, t_2), which only need to contain Top-5 or Top-10 (user-defined) item types and themselves' frequency, then, summary all these bins in a side-by-side way (see figure 2). By this way, synopsis is not only maintain the partial approximate results which have been done, but also remove obsolete data in a updating manner to predict for real-time incoming stream data. It's a very effective method to further reduce space usage in memory and improve the computations. SFI algorithm is also a more adaptive and fast way to mine the most frequent items in specific stream resources. Finally, our experiment results show that our proposed method has a significant performance benefit of improving the accuracy of approximated answers for the real time traffic analyses over packet streaming networks.

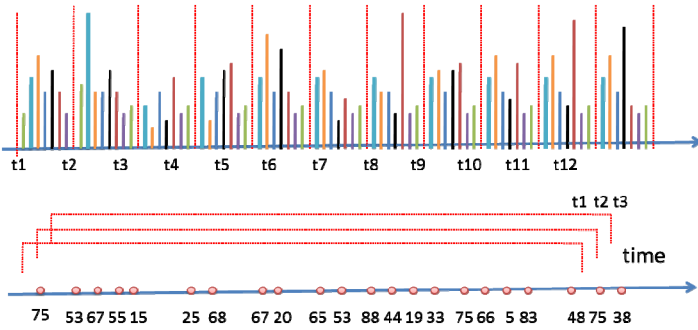


Fig. 2. Synopsis assumption

2 Related Work

Internet traffic on a high-speed link arrives so fast that useful sliding windows for original stream data also would be too large to fit in main memory. In this case, the window must somehow be summarized and an answer must be approximated on the basis of the available synopsis information. When solving the queries over the whole sliding windows, only store a summary in sub-windows which could be resolved by slide manager. Then, queries can be re-evaluated cause by the most recent sub-windows is full. Synopsis is a list expressed by these presented summaries for each sub-window; it is a very important mining method for the stream to compute the statistics rapidly. Many different synopsis data structures have been mentioned in streaming processing [7, 8, 9, and 10].

How to return a list of frequently occurring items is still an important issue, especially for periodically refreshed Top-k queries over sliding windows on internet traffic streams [11]. In the [6], they use a regular grid to index the valid records in main memory. The initial result of each query is computed by a Top-k computation module that processes the minimum number of cells. Only record updates within these cells can potentially invalidate the current Top-k set. Therefore, the maintenance of the result considers only point arrivals and expirations therein. It's a recent work for solving Top-k queries over sliding windows. Another work in [12] considers the problem of efficiently finding the Top-k answers for join queries over web-accessible database. FREQUENT [11], which identifies frequently occurring items in sliding windows and estimates their frequencies. It answer the frequent item queries using small-size basic-window synopses, but no obvious rule for merging the partial information in order to obtain the final answer. James Cheng et al. in [13] proposed a false-negative approach to approximate the set of frequent itemsets (FIs) over a sliding window. Existing approximate algorithms use an error parameter ϵ , to control the accuracy of the mining result. However, the use of ϵ leads to a dilemma. A smaller ϵ gives a more accurate mining result but higher computational complexity, while increasing ϵ degrades the mining accuracy. They address this dilemma by introducing a progressively increasing minimum support function. When an itemset is retained in the window longer, require its minimum support to approach the minimum support of

an FI. Thus, the number of potential FIs to be maintained is greatly reduced. Finding frequent itemsets is the most costly task in association rule mining [14]. Outsourcing this task to a service provider brings several benefits to the data owner such as cost relief and a less commitment to storage and computational resources. For this purpose, they propose and develop an Audit Environment, which consists of a database transformation method and a result verification method. The main component of their audit environment is an artificial itemset planting (AIP) technique. They provided a theoretical foundation on their technique by proving its appropriateness and showing probabilistic guarantees about the correctness of the verification process.

3 Stream Data Processing Model

Since data streams are potentially unbounded in size, the amount of storage required to compute an exact answer to a data stream query may also grow without bound. While external memory algorithms are always used for traditional databases, they are not well suited to data stream applications since they do not support continuous queries and are typically too slow for real-time response. The continuous data stream model is most applicable to problems where timely query responses are important and there are large volumes of data that are being continually produced at a high rate over time. When we are limited to a bounded amount of memory it is not always possible to produce exact answers for data stream queries; however, high-quality approximate answers are often acceptable in lieu of exact answers. Approximation algorithms for problems defined over data stream has been a fruitful research area in the algorithms community in recent years. In our design, we use a slide manager to control the real-time process over multi-streams (as shown in figure 3). There refers to contain three major contents of slide manager. First, slide manager that is to keep tracking of time and determining when to end the next slice. Second, some sub-windows are produced by slide manager decision, and then only store a Top-k summary list in each sub-window. Synopsis is always maintained the whole summary lists and update them by new items arrive. Third, these history data would be removed from the main memory to the secondary storages when slide manager decides the whole sliding windows stopped.

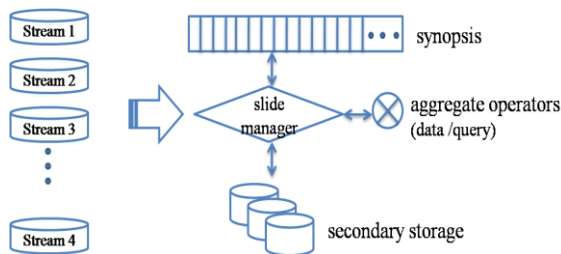


Fig. 3. Stream data processing

3.1 Motivation

Sometimes, all of the sub-windows have the same or similar sizes such as basic-window synopsis. It is very easy to implement, but it's very hard to share the overlapping windows when solving the multi-aggregate queries over the whole streams, because it always leads to more slices. Some unusual synopsis (paired-window) can solve this problem well; it's very efficient to process the multi-aggregate queries over sharing the overlapping windows [15]. The advantage of synopsis that reduce both space and computation cost of evaluating sliding window queries by sub-aggregate and sharing computation. However, for identifying the frequent items over unusual-synopsis, the problem is that false negatives always occur very highly for the Top-k process (as shown in figure 4). Therefore, if the sub-window is fixed, a big count difference among frequencies is bad for evaluating the frequent items. Otherwise, if the sub-window is not fixed, it's very difficult to retrieve the Top-k list. So, in this paper, we want to solve two important problems. The first problem is that how to reduce the risk of sudden rising frequencies in a special time interval. And the second problem is that how to track the real-time frequency for each packet stream, especially for unfixed windows.

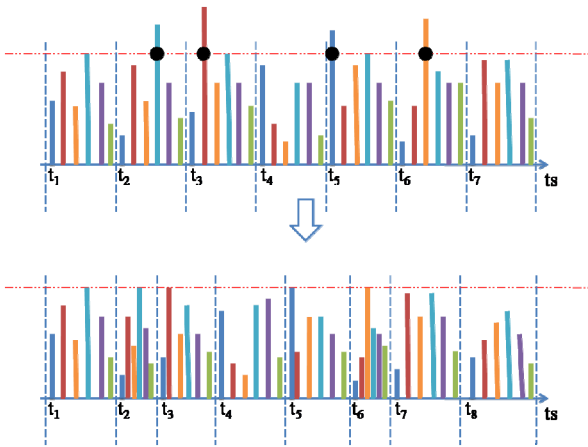


Fig. 4. Smooth synopsis transformation

4 Real-Time Tracking Top-k Query

Normally, people are always count the Top-k items after dividing the time intervals in advance. The way is easy that leads to the first problem, our proposed SFI algorithm dedicated to solve the instability of the results because of a huge different frequency. It through by controlling the top frequency count to restructure synopsis, in order to improve the stability of the Top-k operation. SFI is also a stream Top-k algorithm which could fast and adaptive to count the most frequent item, furthermore to obtain a safe relative result by using a smooth synopsis. As shown in figure 5, there are several

unfixed sub-windows are generated after restructured synopsis. Unlike the traditional way, our SFI algorithm is focus on the top-1 items variation which could be judged by the similarity degree after the previous numerical and a new coming numerical value comparison in a sorting list. Then, a corresponding frequent items list is automatically generated. When the maximum reach a predetermined value, this automatically generated list would be pointed out a list of frequencies for each item. By this way, it could efficient to solve the second problem.

4.1 Double Top-k Phases

Actually, some mining techniques can be processed in each sub-window as a preprocess aggregates (such as Top-k), and these sub-results could be combined and reprocessed in synopsis for each sliding windows workload. Based on the results which have been calculated by each sliding window, we can predict the final approximate query result to the users. Here, we consider synopsis that can be stored in the allotted amount of memory and can be used to provide approximate answers to users' queries along with some reasonable guarantees on the quality of the approximation. In this paper, our proposed method is not only a novel Top-k algorithm for stream process, but also is a superposition of distributed approximate process method. For example, in the figure 5, a Top-k list could be calculated for each sub-window and maintain them into a dynamic summary which is always update by new items go into and old items removed. In fact, we use double Top-k phases to do this job in the whole communication process. It's easy to implement upper-top-k for the whole window, because the size of the entire window is a fixed size. The general Top-k algorithm could be used for this phase. However, the lower-top-k algorithm is needed to calculate the most frequent items as soon as possible, especially for the rapid bursts data. The problem is that the size of sub-window is usually not fixed. That's why the general Top-k algorithm is not suit for this kind of problem. SFI is an effective algorithm to do this job.

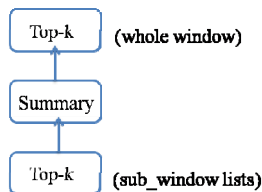


Fig. 5. Double Top-k phases

4.2 SFI Algorithm

SFI algorithm mines the most frequent items by a more adaptive and fast way in specific stream resources. Classify the whole arrived data in memory is still a large work in real time processing, although the result is always changed by time running. So, we propose that to operate the counter in a distributed way. It based on each

personal sub-window to predict the most frequent item for the whole data. It requires more fast operated algorithm and fast response results on the final synopsis. In our opinion, only need to track the sensitive Top-1 item and frequency in order to reconstruct a smooth synopsis.

Input: stream data set (e), a list of sorted tuples ($sort_list$), number of tuples in one type ($count$), number of tuples for each tuple type ($count_list$), large compare function ($large$), the maximum value of frequency (θ).

Output: the maximum frequency count (C_{max}), the maximum frequent item (I_{max}).

Repeat:

1. For increasing stream data e :
 - For sort all tuples e in the initialized sub-window into $sort_list$.
2. For reading the stream data from the $sort_list$:
 - For use large function to compare the vaules of items e ;
 - If $item\ e$ equals to the next one;
 - $count$ adds 1.
 - Otherwise $count$ equals to 0.
3. For reading the stream data from frequent $count_list$:
 - For sorting all the tuples from $count_list$ and initialize the value of C_{max} ;
 - For indexing to I_{max} .
4. Output C_{max} and I_{max} .

5 Experiments and Evaluation

In this part, we orderly present designs for SFI algorithm and use exact stream data to compare with Traditional Manner (TM). We have tested algorithm SFI on Tuple-based windows over TCP traffic data. The trace contains 199 and 1592 distinct source IP addresses for workload 10000 and 80000, which can be treated as distinct item types. And experiment with two values of θ : $\theta=3$ (sub-window initialization size is 80), $\theta=9$ (sub-window initialization size is 400). The size of the Top-k list, k , is varied from 1 to 10. To build a prototype TCP/IP stream data in C++, and stream data running by real time. We imitate the attribute of TCP/IP addresses to make two kinds stream data set, at the same time sign the items types into each TCP/IP addresses. Then, we design a TCP/IP addresses generator client program to control the data flooding by current time.

5.1 Experimental Results

In the SFI algorithm, θ is a very important evaluation criterion. It has an effect on rising speed of identified frequent items. Some new itemsets have not been identified

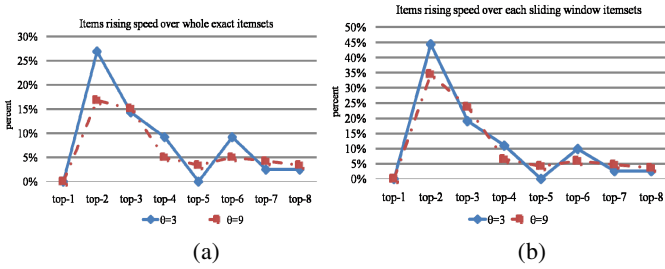


Fig. 6. (a)shows identified frequent itemsets rising speed over whole exact itemsets and (b) over for each sliding window itemsets when variable θ with 3 and 9.

completely when the rising speed is high, it may lead to more data would be ignored. Otherwise, almost all kinds of itemsets would be calculated in the upper-processing. Although it may give us a more exact result, more space usages are needed. Figure 6(a) shows the effect of rising speed rate on the whole itemsets, there more itemsets have been identified when θ equals to 3. It shows a high varied speed between Top-1 and Top-3; even there has a smooth balance with no more new itemsets will come by Top-5. There is a different state with smoothly rising by θ equals to 9. In figure 6(b), we discuss identified frequent itemsets rising speed by variable θ with 3 and 9. There is a small difference with figure 6(a) is almost half of the initialization datasets have been identified over each sliding window by θ equals to 3, and in figure 6(a) the identified datasets only 25% account for the whole datasets. Another point is that similar rising speeds based on their own sliding windows datasets. Generally, the value of parameter θ is too large to get a bad accuracy, because it can affect the size of the sub-window. Otherwise, the value of parameter θ is too small to get a high space usage. In the testing, the accuracy of identified frequent items is raising by θ regression.

Next, in order to test the result of Top-11 streaming items over the whole datasets based on sub-window equals to 80 and 400 itemsets. As shows in figure 7, when Top-1, then 10 items can be identified by sub-window (80). There are all of Top-11 steaming items can be identified from Top-2. Compare with sub-window equals to 400 items, Top-1 only can identify 8 items and there can be identified all the datasets with contained the Top-11 query items from Top-5. Therefore, how to choose an appropriate sub-windows size is very important. Here, sub-window equals to 80 is more better than 400.

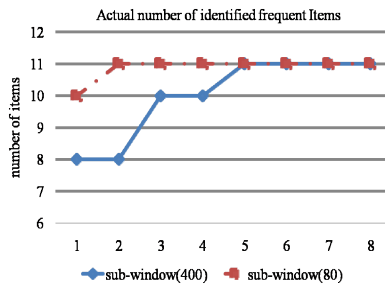


Fig. 7. Actual number of identified frequent items over sub-window different size with 80 and 400 items

5.2 Space Usage

SFI algorithm implements a simple count operator to reduce space usage: randomly sampling items to be stored in the synopses and deleting parts of older synopses over sliding windows. The sliding window sizes in our experiments are 10000 and 80000 , which may be considered as a rough estimate for the space usage. The space usage of our algorithm is significantly smaller, especially when sub-window size is large and/or k is small as shown in figure 8. Because a top-k sketch must be stored for each sub-window, the number of sub-windows has the greatest effect on the space requirements. Another important one is that we designed a simple count algorithm SFI instead of traditional counter (TM) process in real time analyses in order to reduce highly memory space usage and computation time (see figure 9).

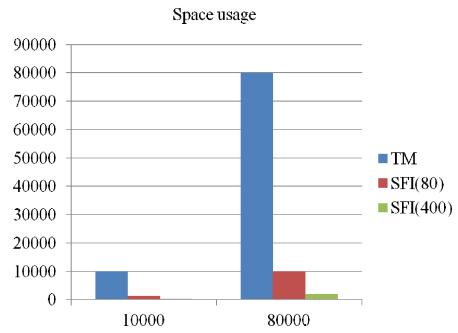
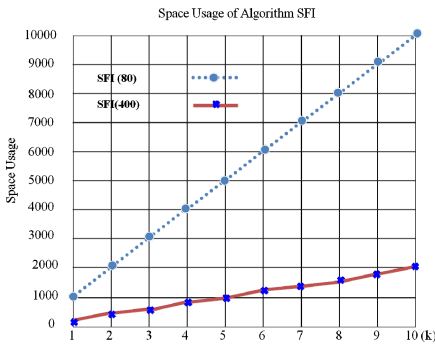


Fig. 8. Space usages of SFI algorithm for different sizes of sub-windows **Fig. 9.** Space usages of SFI algorithm and TM

6 Conclusion

In this paper, we proposed SFI algorithm for Top-k query to mine the most frequent items by a more adaptive and fast way in specific stream resources. It focuses on the Top-1 items variation which could be judged by the similarity degree after the previous numerical and a new coming numerical value comparison in a sorting list. Then, a corresponding frequent items list is automatically generated. Actually, it's not only a novel Top-k algorithm for stream process, but also is a superposition of distributed approximate process method (double-top-k process). Finally, our experiment results show that using SFI has a significant performance benefit of improving the accuracy of approximated answers for the real time traffic analyses over packet streaming networks. In future work, we intend to solve more complex integration of different kinds of synopses for the multi-aggregate queries. Furthermore, we want to extend to support real time mining and classification using synopses.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.51077010) and by the Jilin provincial department of science and technology (No.20120338).

References

1. Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Nishizawa, I., Rosenstein, J., Widom, J.: STREAM: The Stanford Stream Data Manager (Demonstration Description). In: 2003 ACM SIGMOD International Conference on Management of Data, p. 665 (2003)
2. Li, J., Tufte, K., Shkapenyuk, V., Papadimos, V., Johnson, T., Maier, D.: Out-of-order Processing A New Architecture for High-Performance Stream Systems. In: 34th International Conference on Very Large Data Bases, VLDB 2008, pp. 274–288 (2008)
3. Chandrasekaran, S., Cooper, O., Deshpande, A., Franklin, M.J., Hellerstein, J.M., Hong, W., Krishnamurthy, S., Madden, S.R., Reiss, F., Shah, M.A.: TelegraphCQ: Continuous Dataflow Processing for an Uncertain World. In: First Biennial Conference on Innovative Data Systems Research, CIDR 2003, Asilomar (2003)
4. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models And Issues in Data Streams. In: 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database System, pp. 1–16 (2002)
5. Pavlo, A., Paulson, E., Rasin, A., Abadi, D.J., Dewitt, D.J., Madden, S., Stonebraker, M.: A Comparison of Approaches to Large-Scale Data Analysis. In: 2009 ACM SIGMOD International Conference on Management of Data, pp. 165–178 (2009)
6. Mouratidis, K., Bakiras, S., Papadias, D.: Continuous Monitoring of Top-K Queries over Sliding Windows. In: 2006 ACM SIGMOD International Conference on Management of Data, pp. 635–646 (2006)
7. Krishnamurthy, S., Wu, C., Franklin, M.J.: On-the-fly Sharing for Streamed Aggregation. In: 2006 ACM SIGMOD International Conference on Management of Data, pp. 623–634 (2006)
8. Toman, D.: On Construction of Holistic Synopses under the Duplicate Semantics of Streaming Queries. In: TIME 2007 14th International Symposium on Temporal Representation and Reasoning, pp. 150–162 (2007)
9. Stern, M., Buchmann, E., Bohm, K.: A Wavelet Transform for Efficient Consolidation of Sensor Relations with Quality Guarantees. In: VLDB 2009, 35th International Conference on Very Large Data Bases, pp. 157–168 (2009)
10. Matias, Y., Urieli, D.: Optimal Workload-Based Weighted Wavelet Synopses. *Journal of Theoretical Computer Science*, 227–246 (2007)
11. Golab, L., DeHaan, D., Demaine, E.D., Lopez-Ortiz, A., Munro, J.I.: Identifying Frequent Items In Sliding Windows Over On-Line Packet Streams. In: 3rd ACM SIGCOMM Conference on Internet Measurement, pp. 173–178 (2003)
12. Wu, M., Equille, L.B., Marian, A., Procopiuc, C.M., Srivastava, D.: Processing Top-k join Queries. In: VLDB 2010, 36th International Conference on Very Large Data Bases, pp. 860–870 (2010)
13. Cheng, J., Ke, Y.: Maintaining Frequent Itemsets over High-Speed Data Streams. In: 2006 PAKDD, 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 462e–467e (2006)
14. Wong, W.K., Cheung, D.W., Hung, E., Kao, B., Mamoulis, N.: An Audit Environment for Outsourcing of Frequent Itemset Mining. In: VLDB 2009, 35th International Conference on Very Large Data Bases, pp. 1162–1172 (2009)
15. Wang, L., Koo Lee, Y., Ryu, K.H.: Supporting Top-K Aggregate Queries over Unequal Synopsis on Internet Traffic Streams. In: Zhang, Y., Yu, G., Bertino, E., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 590–600. Springer, Heidelberg (2008)

Sequential Diagnosis Method for Rotating Machinery Using Support Vector Machines and Possibility Theory

Hongtao Xue¹, Ke Li², Huaqing Wang³, and Peng Chen^{1,*}

¹ Graduate School of Bioresources, Mie University, 514-8507, Japan
xueht@live.jp, chen@bio.mie-u.ac.jp

² School of Mechanical Engineering, Jiangnan University, Wuxi, 214122, China
dayanlv@live.cn

³ Beijing University of Chemical Technology, Beijing, 100029, China
wanghq_buct@hotmail.com

Abstract. This study proposes a new intelligent diagnosis method for rotating machinery using support vector machines (SVMs) and possibility theory on basis of the vibration signals, to detect faults and identify fault types at an early stage. The non-dimensional symptom parameters (NSPs) in the frequency domain are defined to reflect the features of the vibration signals. SVMs are used to construct the optimal classification hyper-plane, then to integrate a new symptom parameter, which is called as synthetic symptom parameter (SSP). Finally, the possibility distributions of the SSP are used to distinguish faults by sequential inference and possibility theory. The proposed method has been applied to detect the faults of the V-belt driving equipment in a centrifugal fan, and the efficiency of the method has been verified using practical examples.

Keywords: Sequential diagnosis, support vector machine, possibility theory, synthetic symptom parameter, V-belt driving equipment.

1 Introduction

Vibration feature is the most revealing information on the condition of rotating machinery [1]. Therefore, vibration signal is often used for fault detection and state discrimination in rotating machinery in the field of on-line monitoring and condition diagnosis of plant machinery. However, there are some problems in the condition diagnosis using vibration signal. Firstly, feature extraction for fault diagnosis is difficult because when the vibration signal for the condition diagnosis is measured at a position far from the faulty part or a fault is at an early stage, the signal measured for the condition diagnosis may contain strong noise. Then it is not easy to find effective symptom parameters for detecting faults and identifying fault types, and the detection sensitivity of SPs is not high [2]. Secondly, there are many ambiguous relationships between the symptom parameters and the failure types. Finally, some traditional intelligent methods such as neural networks, which are based on statistical learning, require a large number of training samples. However, in most cases of practical machinery diagnosis, measured signals are not sufficient for the learning data, and the

* Corresponding author.

learning of neural network often does not converge due to the infelicity learning data. For the above reasons, this study proposes a new intelligent diagnosis method for rotating machinery using support vector machines (SVMs) and possibility theory, to detect faults and identify fault types at an early stage by using the vibration signals. Fig. 1 shows the flow diagram for the proposed procedure. These proposed methods have been applied to diagnose the faults of the V-belt driving equipment in a centrifugal fan, and verification results will show that the methods are accurate and practical.

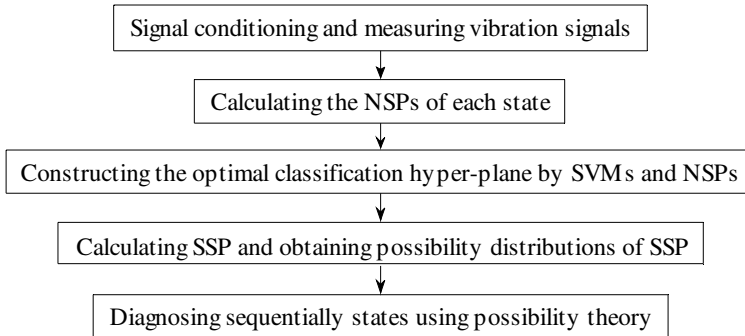


Fig. 1. Flow chart of fault diagnosis system

2 V-Belt Driving Equipment for Experimental Verification

The V-belt driving equipment used in the centrifugal fan for condition diagnosis is shown in Fig. 2. The motor is employed to drive the centrifugal fan through 2 V-belt, and the rotation speed can be varied through a speed controller. Three accelerometers are mounted on the three directions (horizontal, vertical and axial direction) of the pump housing to measure the vibration signals with a sampling frequency of 50 kHz. Aiming at some frequent faults of driving equipment, such as pulley defect (PD), belt defect (BD) and belt looseness (BL) states, those faults and normal state (N) have been set to measure the vibration signals.



Fig. 2. The centrifugal fan using V-belt driving for condition diagnosis

3 Non-dimensional Symptom Parameters

In order to diagnose intelligently the condition of plant machinery, symptom parameters (SPs) are required to express the information indicated by a signal measured for diagnosing machinery faults. And a good symptom parameter can correctly reflect states and the condition trend of plant machinery [3]-[4]. A large set of symptom parameters has been defined in the pattern recognition field. Here, analyzing the spectral features of the faults of the V-belt driving equipment, 3 non-dimensional symptom parameters (NSPs) in frequency domain, commonly used for the condition diagnosis of V-belt driving equipment, are defined.

$$P_1 = \sqrt{\frac{\sum_{i=1}^I f_i^2 \cdot F(f_i)}{\sum_{i=1}^I F(f_i)}} \quad (1)$$

$$P_2 = \frac{\sum_{i=1}^I (f_i - \bar{f})^3 \cdot F(f_i)}{\sigma^3 \cdot I} \quad (2)$$

$$P_3 = \frac{\sum_{i=1}^I (f_i - \bar{f})^4 \cdot F(f_i)}{\sigma^4 \cdot I} \quad (3)$$

where I is the number of spectrum lines, f_i is the frequency from 0 Hz to the maximum analysis frequency, $F(f_i)$ is the power spectrum value of f_i , \bar{f} is the mean value of the frequency, such that $\bar{f} = \sum_{i=1}^I (f_i \cdot F(f_i)) / \sum_{i=1}^I F(f_i)$ and σ is the standard deviation and $\sigma = \sqrt{\sum_{i=1}^I ((f_i - \bar{f})^2 \cdot F(f_i)) / I}$.

Popularly, the sensitivity of a SP is judged by discrimination index (DI) [5] as follows. And then it is proved that the larger the value of the DI, the higher the sensitivity of the SP will be, and therefore, the better the SP will be.

$$DI = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (4)$$

where μ_1 , σ_1 and μ_2 , σ_2 are the mean and standard deviation of the SP in two states.

However, it is not easy to find an effective SP for some faults. Moreover, the DI values of these SPs are almost lower at an incipient stage of a fault. Therefore, in the study, synthetic symptom parameter (SSP) is defined with the function of the optimal classification hyper-plane with primitive NSPs using soft SVM.

4 Support Vector Machine (SVM)

SVM is a relatively new computational learning method based on the statistical learning theory. The goal of SVM is to define an optimal classification hyper-plane that

divides all samples so that the samples with the same label are on the same side of the classification hyper-plane, and the distance between two classes and the classification hyper-plane is maximized [6]-[8]. In the case of linear data, a classification hyper-plane can be expressed as follows.

$$\omega \cdot x + b = 0 \tag{5}$$

where, ω is N-dimensional vector, x is the input data, b is a classification threshold.

However, the cases of linear separation are rare on the real problem of pattern recognition. But then if minimum error is permitted, and the condition for the optimal classification hyper-plane is relaxed, the optimal classification hyper-plane using linear SVM can be constructed. This is soft margin SVM, as shown in Fig. 3, in a 2-dimensional situation.

Of course, for a linear non-separable case, SVM can perform a non-linear mapping of the input vector x from the input space into a higher dimensional space, where the mapping is determined by kernel function. But the vibration signals from different conditions have only changing tendency, but not obvious boundary.

Then in the field of mechanical fault diagnosis, soft margin SVM has a strong generalization capacity when only a small amount of training samples are available. These features have an important role in the field of mechanical fault diagnosis.

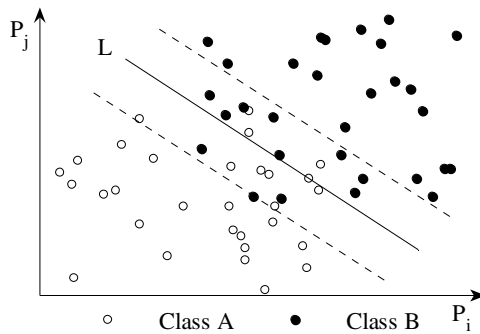


Fig. 3. Example of the optimal classification hyper-plane constructed by soft margin SVM

5 Defining Synthetic Symptom Parameter by SVM

Soft margin SVM is used to construct the optimal classification hyper-plane, and the function corresponding to the optimal classification hyper-plane is called “the optimal classification function” as follows.

$$f(x) = \omega \cdot x + b \tag{6}$$

In this study, in order to increase the diagnostic sensitivity of the SPs, the optimal classification function obtained using soft margin SVM and primitive SPs is defined as synthetic symptom parameter (SSP).

$$SSP = \omega \cdot [P_1 \ P_2 \ \dots \ P_n] + b \tag{7}$$

To explain the efficiency of SSP defined using soft margin SVM and primitive SPs, two SPs are used to obtain the optimal classification hyper-line (L) when class A and class B are distinguished. Then a line perpendicular to the line (L) is SSP-axis, as shown in Fig. 4. $f_A(P_i)$ and $f_B(P_i)$ are the P_i distributions of class A and class B, $f_A(P_j)$ and $f_B(P_j)$ are the P_j distributions of class A and class B, $f_A(SSP)$ and $f_B(SSP)$ are the SSP distributions of class A and class B. It is obvious that the overlap of $f_A(SSP)$ and $f_B(SSP)$ is smaller than those of others. Then the sensitivity of SSP is higher than P_i and P_j .

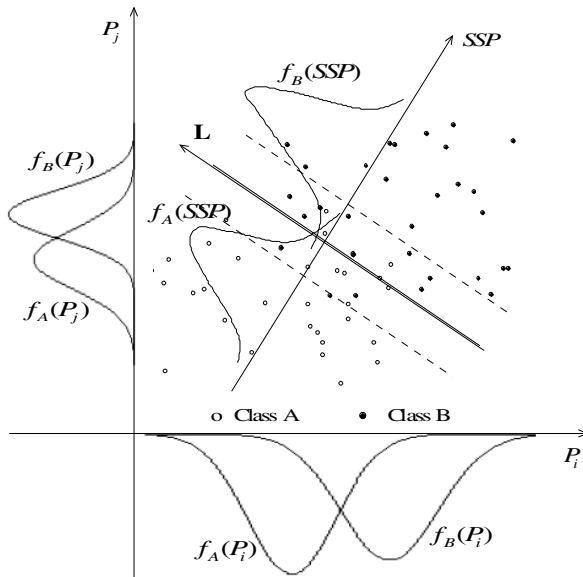


Fig. 4. The efficiency of SSP

6 Sequential Diagnosis Using Possibility Theory

6.1 Possibility Theory

For fuzzy inference, the membership function of a SP is necessary [9]-[10]. This can be obtained from probability density functions of the SP using possibility theory. When a SP conforms to the Weibull distribution, it can be changed to possibility distribution function $p(x_i)$ by the following formulae.

$$p(x_i) = \sum_{k=1}^N \min\{\lambda_i, \lambda_k\}, \quad (i, k = 1, 2, \dots, N) \tag{8}$$

λ_i and λ_k can be calculated as follows:

$$\lambda_i = \int_{x_{i-1}}^{x_i} \frac{m}{\eta} \left(\frac{x-x_0}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x-x_0}{\eta}\right)^m\right\} dx, x_i = \bar{x} + \frac{6i-3N}{N} \cdot \sigma \tag{9}$$

$$\lambda_k = \int_{x_{k-1}}^{x_k} \frac{m}{\eta} \left(\frac{x-x_0}{\eta}\right)^{m-1} \exp\left\{-\left(\frac{x-x_0}{\eta}\right)^m\right\} dx, x_k = \bar{x} + \frac{6k-3N}{N} \cdot \sigma \tag{10}$$

Where, N is the division number of the domain of the SP namely $[\bar{x}-3\sigma, \bar{x}+3\sigma]$, \bar{x} and σ are the mean and the standard deviation of the SP, respectively. m , η and x_0 are the parameters of the shape, scale and location. Fig. 5 (A) shows an illustration of the possibility function and the probability density function.

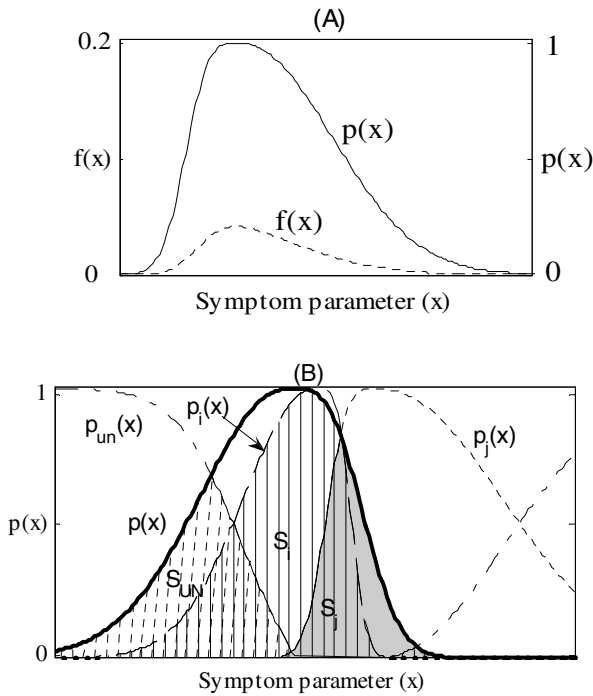


Fig. 5. (A) Probability density function $f(x)$ and possibility function $p(x)$ based on Weibull distribution; (B) Matching examples of possibility function based on Weibull distribution;

When intelligently identifying a state, the possibility function of the SP is used to constitute the membership function for diagnosis using possibility theory. Fig. 5 (B) shows the matching examples of possibility function. The common area, between the possibility functions $p_i(x), p_j(x)$ of model state i, j and the possibility functions $p(x)$ of diagnostic state, is calculated by the following formulas.

$$S_i = \int_{\bar{x}_d - 3\sigma_d}^{\bar{x}_d + 3\sigma_d} \psi_i(x) dx, \quad \psi_i(x) = \min\{p_i(x), p(x)\} \quad (11)$$

$$S_j = \int_{\bar{x}_d - 3\sigma_d}^{\bar{x}_d + 3\sigma_d} \psi_j(x) dx, \quad \psi_j(x) = \min\{p_j(x), p(x)\} \quad (12)$$

Where \bar{x}_d and σ_d are the mean and the standard deviation of the SP in diagnostic state. Moreover, the unknown state except state i, j is the model state un , its possibility functions $p_{un}(x)$ is calculated by the following formula (13). And the common area between $p_{un}(x)$ and $p(x)$ is calculated by the following formula (14).

$$p_{un}(x) = \max\{1 - [p_i(x) + p_j(x)], 0\} \quad (13)$$

$$S_{un} = \int_{\bar{x}_d - 3\sigma_d}^{\bar{x}_d + 3\sigma_d} \psi_{un}(x) dx, \quad \psi_{un}(x) = \min\{p_{un}(x), p(x)\} \quad (14)$$

Therefore, the probability, a diagnostic state should be diagnosed with model state i, j and un , respectively, can be calculated by the following formulas. Then it is judged that the diagnostic state is the model state with the most probability.

$$\omega_i = S_i / (S_i + S_j + S_{un}) \quad (15)$$

$$\omega_j = S_j / (S_i + S_j + S_{un}) \quad (16)$$

$$\omega_{un} = S_{un} / (S_i + S_j + S_{un}) \quad (17)$$

6.2 Building Diagnostic System

SSP, as the object of a follow-on process, is obtained using soft margin SVM and primitive 3 NSPs and is used for condition diagnosis. However, SVM is a two-class classifier, and there are many and complex fault types which may occur in a rotating machine. In order to achieve an intelligent and efficient condition diagnosis, sequential diagnosis is proposed, as shown in Fig. 6. Then the diagnostic system is designed to consist of a number of possibility functions of SSPs and is able to diagnose the faults sequentially. The first step is used to detect faults, and the other steps are used to identify the fault type. In order to improve performance, it is recommended that conditions that represent more frequent faults of V-belt driving equipment should be placed in the diagnostic system. Moreover, V-belt faults consist of pulley defect, belt defect and belt looseness in the first step, and other V-belt faults consist of belt defect and belt looseness in the second step. Therefore, the training data of each step are input SVM to obtain the SSP for diagnosis as shown in Table 1. Then the possibility functions of SSPs in each step are used to build the diagnostic system.

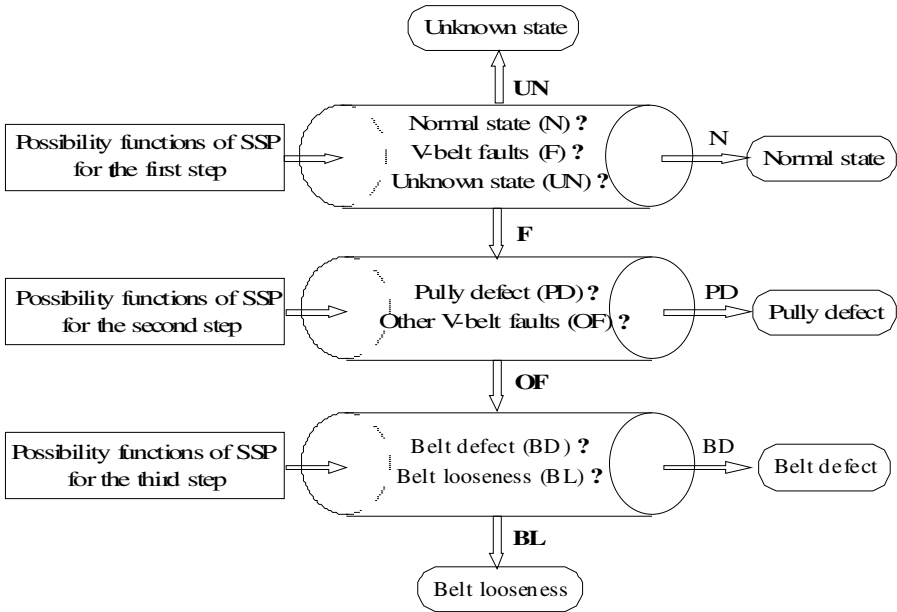


Fig. 6. Flowchart of the sequential diagnosis

Table 1. Diagnosis goal and the SSPs of each step

step	Diagnosis goal	synthetic symptom parameter (SSP)
1	Normal state	$y = 61.9P_1 - 15.4P_2 + 10.9P_3 - 33.7$
2	Pulley defect	$y = -35.2P_1 + 56.8P_2 - 40.1P_3 + 17.4$
3	Belt defect or belt looseness	$y = 98.1P_1 + 9.25P_2 - 46.36P_3 - 24.3$

When one state will be diagnosed, the primitive 3 NSPs are calculated to input the diagnostic system. Firstly, the SSPs in the first step are calculated to acquire the possibility function of the SSPs. Secondly, the probabilities diagnosed with model states (normal, V-belt faults and unknown state) can be calculated by the formulas (11)-(17). Finally, the system judges intelligently by the probabilities of model states. If the diagnostic state is judged normal state, the system can stop. Otherwise, the diagnostic state is judged V-belt faults, and the second step will be performed to judge whether a fault is pulley defect or not. More faults can be judged in the same manner.

6.3 Diagnosis Classification

To verify the proposed diagnostic method, the test data, measured in each known state and not used for the pre-calculated possibility function, are used. The practical diagnostic example for diagnosing one state is shown in Fig. 7. In the first step, the probabilities of the state diagnosed with normal, V-belt faults and unknown state are 0.192,

0.577, 0.231, respectively. Therefore, it is judged that one V-belt fault has occurred, and then the second step is performed. The probability of pulley defect is 0.741, it is obvious that the state is pulley defect. Similarly, four test states (N, PD, BD, BL) have been identified correctly, the diagnosis results are shown in Table 2.

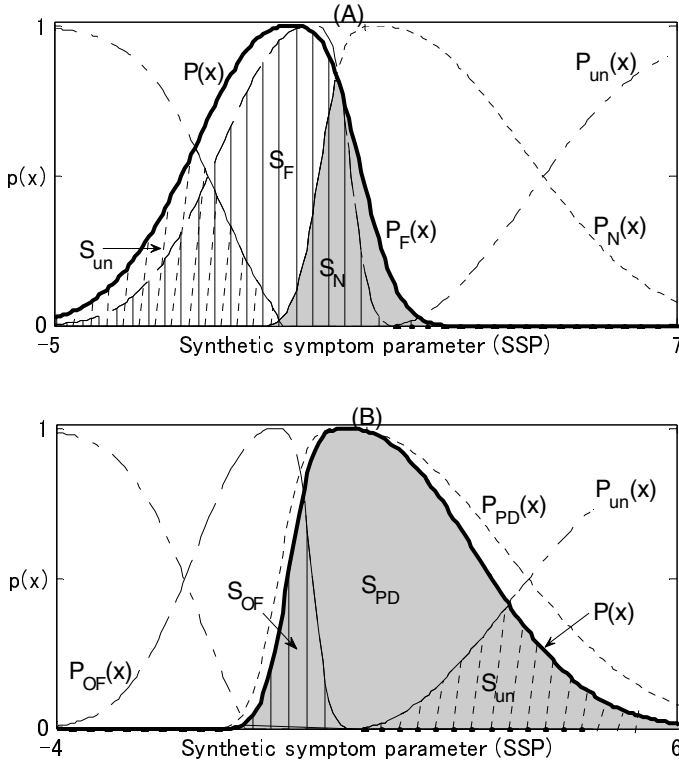


Fig. 7. The practical diagnostic example for diagnosing one state: (A) diagnosis in the first step, (B) diagnosis in the second step.

Table 2. The Diagnostic results from various states

Test state	The probability of each state in each step									Judge state
	In the first step			In the second step			In the third step			
	N	F	UN	PD	OF	UN	BD	BL	UN	
N	0.692	0.201	0.107							N
PD	0.192	0.577	0.231	0.741	0.103	0.156				PD
BD	0.247	0.616	0.137	0.130	0.789	0.081	0.803	0.154	0.043	BD
BL	0.267	0.593	0.140	0.224	0.675	0.101	0.198	0.746	0.056	BL

7 Conclusion

To intelligently diagnose the faults of the V-belt driving equipment at an early stage, a new intelligent diagnosis method using SVMs and possibility theory on basis of the vibration signals was proposed, and the effectiveness was proved according to the experiment. The superiority of the method proposed in this paper can be explained as: (1) SSP defined using SVM and primitive SPs cannot expend time and effort to find effective primitive SPs, and increases the diagnostic sensitivity. (2) The method proposed here has a strong adaptable ability to deal with the faults at an early stage.

The method proposed in this paper will be applied to condition diagnosis in various types of rotating machinery in a real plant.

References

1. Pratesh Jayaswal, A.K.: Machine fault signature analysis. *Int. J. of Rotating Machinery*, Article ID 583982, 10 pages (2008)
2. Chen, P.: *Foundation and Application of Condition Diagnosis Technology for Rotating Machinery*. Sankeisha Press, Japan (2009)
3. Richardson, J.J.: *Artificial Intelligence in Maintenance*. Noyes Publications (1985)
4. Matuyama, H.: Diagnosis Algorithm. *J. of JSPE* 75, 35–37 (1991)
5. Chen, P., Toyota, T., He, Z.: Automated function generation of symptom parameters and application to fault diagnosis of machinery under variable operating conditions. *IEEE Trans. on Syst., Man and Cybernetics, Part A: Systems and Humans* 6, 775–781 (2001)
6. Achmad, W., Yang, B.S.: Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Science Direct*, 241–250 (2007)
7. Yong, D.D., Nam, G., Busan: Support vector machine in machine condition monitoring and fault diagnosis. *Dec.*, 608–739 (2006)
8. Gunn, S.R.: *Support Vector Machines for Classification and Regression* (May 10, 1998)
9. Dubois, D., Prade, H.: *Possibility Theory-An Approach to Computerized Processing*. Plenum Press, New York (1988)
10. Chen, P., Toyota, T.: Sequential fuzzy diagnosis for plant machinery. *JSME Int. J. C* 46(3), 1121–1129 (2003)

On M-Type Bag Structures

Kankana Chakrabarty

Discipline of Computer Science
School of Science and Technology
University of New England, Armidale 2351
New South Wales, Australia
kankanac@turing.une.edu.au

Abstract. In this paper, the author introduces a structure called M-type bag structure which can be defined on a non-empty set associated with an indiscernibility relation. It can be observed that an M-type bag structure represents a bag if the indiscernibility relation be defined in such a way that any two elements of the set are indiscernible under a given set of criteria that considers the values of some predefined attribute set. This paper further studies some algebraic properties of M-type bag structures.

Keywords: Bags Fuzzy Bags; M-type bag structures.

1 Introduction

The notion of bags [2, 4, 9, 10, 14] support knowledge representation associated with the type of domains that allows repeatability of elements in a construct. Bag-projection and bag-selection operations are extremely useful in case of database modelling. Bags are often used as type constructors for modelling purposes when data sub-languages require multiset semantics [1, 5, 7]. The dependencies between attribute sets can be modelled by using the bag and fuzzy bag structures [3, 5, 6, 8, 11, 12, 13, 14]. The indiscernibility of objects plays a crucial role in these cases. Some information systems allow the indiscernibility relation to be defined based on the available knowledge regarding the attribute types of the concerned objects.

This paper introduces a notion called an M-type bag structure that can be defined on a non-empty set on which an indiscernibility relation is defined. It is interesting to note that an M-type bag structure actually represents a bag structure if the concerned indiscernibility relation is defined in such a way that any two elements of the set are indiscernible under a given set of criteria that considers the values of some predefined attribute set. This paper further studies the characteristics of M-type bag structures by investigating some of their algebraic properties.

2 Bags, the Preliminaries

This section briefly presents the notion of bag and its properties [10, 4, 2, 9, 7] that are essential for the development of the concept of M-type bag structures.

A bag β drawn from a set X is represented by a function C_β defined as

$$C_\beta : X \rightarrow \mathbb{N}$$

where \mathbb{N} represents the set of non-negative integers.

In this case, $C_\beta(x)$ represents the number of occurrences of the element x in the bag β . We represent the bag β drawn from the set $X = \{x_1, x_2, \dots, x_m\}$ as

$$\beta = \{x_1/n_1, x_2/n_2, \dots, x_m/n_m\}$$

where n_i is the number of occurrence of the element x_i ($i = 1, 2, 3, \dots, m$) in the bag β .

For any two bags β_1 and β_2 drawn from a set X , the following are defined $\forall x \in X$:

- $\beta_1 = \beta_2$ if $C_{\beta_1}(x) = C_{\beta_2}(x)$
- $\beta_1 \sqsubseteq \beta_2$, i.e β_1 is a sub bag of β_2 , if $C_{\beta_1}(x) \leq C_{\beta_2}(x)$
- $\beta = \beta_1 \oplus \beta_2$ if $C_\beta(x) = C_{\beta_1}(x) + C_{\beta_2}(x)$
- $\beta = \beta_1 \ominus \beta_2$ if $C_\beta(x) = \max[C_{\beta_1}(x) - C_{\beta_2}(x), 0]$

where \oplus and \ominus represents the ‘bag addition’ and ‘bag removal’ operations respectively.

If β be a bag drawn from a set X , then the support set of β denoted as $S(\beta)$ is a subset of X with the characteristic function given by

$$\phi_{S(\beta)}(x) = \min[C_\beta(x), 1]$$

for all $x \in X$. A bag β is called an empty bag if for all $x \in X$, $C_\beta(x) = 0$.

The cardinality of a bag β drawn from a set X is denoted by $|\beta|$ and is defined as

$$|\beta| = \sum_{x \in X} C_{\beta(x)}$$

For a bag β drawn from a set X , $\max_{x \in X} C_\beta(x)$ is said to be the peak value of the bag.

Any $x^* \in X$ satisfying $C_\beta(x^*) = \max_{x \in X} C_\beta(x)$ is called a peak element of the bag β .

The union of two bags β_1 and β_2 drawn from a set X is a bag denoted by $\beta_1 \cup \beta_2$ such that for all $x \in X$, $C_{\beta_1 \cup \beta_2}(x) = \max[C_{\beta_1}(x), C_{\beta_2}(x)]$.

The intersection of β_1 and β_2 results in a bag denoted by $\beta_1 \cap \beta_2$ such that for all $x \in X$, $C_{\beta_1 \cap \beta_2}(x) = \min[C_{\beta_1}(x), C_{\beta_2}(x)]$.

In [4], the Cartesian product of two bags is defined as follows:

For two bags α and β drawn from the sets X and Y respectively, their Cartesian product denoted by $\alpha \otimes \beta$ is a bag drawn from $X \times Y$ such that for all $(x, y) \in (X \times Y)$,

$$C_{\alpha \otimes \beta}(x, y) = C_\alpha(x).C_\beta(y).$$

3 M-Type Bag Structures

This section introduces the notion of M-type bag structures and presents some of their characteristics.

Let X be a non-empty set and let R be an indiscernibility relation defined on X such that under R , X can be decomposed into n partitions P_1, P_2, \dots, P_n . That is,

$$X = \bigcup_{i=1}^n P_i.$$

Then, an M-type bag structure B_X^R drawn from X is characterized by a function $\phi_{B_X^R}$ defined on X such that $\phi_{B_X^R} : X \rightarrow N$, N being the set of non-negative integers, such that for each $x \in X$, $\phi_{B_X^R}(x) = |[x]_R|$, where $|[x]_R|$ represents the cardinality of the equivalence class of x under R .

Clearly, a comparison of the bag structure to that of the M-type bag structure indicates that a crisp bag is nothing but a special case of the M-type bag structure. This is due to the fact that the relation of equality can be considered as a specific case of the concerned indiscernibility relation.

The M-type bag structure B_X^R drawn from X can be represented as $B_X^R = \{(x_i, \phi_{B_X^R}(x)) : x_i \in X\}$.

It should be noted that an M-type bag structure B_X^R will represent a bag if R be defined in such a way that for any $x_1, x_2 \in X$, $x_1 R x_2$ holds if and only if x_1 and x_2 are indiscernible under a given set of criteria that considers the values of some pre-defined attribute set.

The application of the notion of M-type bag structure can help to model as well as manage the domain knowledge gathered from the experts in an effective and efficient way.

As an example, we consider a set E consisting of n experts with expertise in any specific area of domain knowledge. We also consider R be an indiscernibility relation defined on E that decomposes it into m disjoint equivalence classes. Let R be defined in such a way that for any E_1, E_2 in E , E_1 and E_2 belongs to the same equivalence class under R if and only if they have the same level of expertise judged on the basis of p predefined criteria. In this case, the M-type bag structure B_E^R will support knowledge representation that involves the granularity of the object collections in respective partitions P_1, P_2, \dots, P_m formed under R .

Lemma 1. For any non-empty set X , if R represents an equality relation defined on R , then any M-type bag structure B_X^R drawn from X represents a subset of X .

In this case, each partition will consist of just one element of X .

Lemma 2. If B_X^R be any M-type bag structure drawn from a non-empty set X, then for all

$$x \in B_X^R, \phi_{B_X^R}(x) > 0.$$

This is clearly related to the fact that each partition of X under R is non-empty.

Lemma 3. For any non-empty set X , if R represents an equality relation defined on R, then for any M-type bag structure B_X^R drawn from X , $S(B_X^R)$ equals B_X^R .

This presents an inter-relationship of bag structures and M-type bag structures.

Lemma 4. If B_X^R be any M-type bag structure drawn from a non-empty set X, then $S(B_X^R) \neq \emptyset$.

In this case, atleast one partition of X under R will exist as X in non-empty.

Proposition 1. For any M-type bag structure B_X^R drawn from a non-empty set X ,

$$|S(B_X^R)| = |B_X^R|.$$

This presents a clear relationship between an M-type bag structure and its support set.

Proposition 2. If for any non-empty set X , X_1 and X_2 be two subsets of X such that $X_1 \cup X_2 = X$, then for any M-type bag structure $B_{X_1}^R$ drawn from X_1 and for any M-type bag structure $B_{X_2}^R$ drawn from X_2

$$|S(B_{X_1}^R) \cup S(B_{X_2}^R)| = |X|.$$

This demonstrates the relationship between a non-empty set and a pair of M-type bag structures drawn from it.

Proposition 3. For any M-type bag structure B_X^R drawn from a non-empty set X , if any two elements $x_1, x_2 \in X$ are such that they belong to the same partition P_i of X, then

$$\phi_{B_X^R}(x_1) = \phi_{B_X^R}(x_2).$$

This presents a clear relationship between the objects of the same partition under R.

Proposition 4. If B_X^R be any M-type bag structure drawn from a non-empty set X that can be decomposed into n number of partitions under R, then the cardinality of $S(B_X^R)$ is equal to n.

If B_X^R be any M-type bag structure drawn from a non-empty set X and if $Y \subseteq S(B_X^R)$, then B_Y^R is called an M-type sub bag structure of B_X^R if it can be characterized by a function $\phi_{B_Y^R}$ defined on Y such that $\phi_{B_Y^R} : Y \rightarrow N$ where for each $y \in Y$, $\phi_{B_Y^R}(y) = \phi_{B_X^R}(y)$.

If B_X^R be any M-type bag structure drawn from a non-empty set X and if B_Y^R is an M-type sub bag structure of B_X^R , then B_Y^R is itself an M-type bag structure drawn from X .

Lemma 5. If for any non-empty set X , $B_{X_1}^R$ and $B_{X_2}^R$ be any two M-type bag structures drawn from X , then for all $x \in (B_{X_1}^R \circ B_{X_2}^R)$ where \circ represents ‘ \cup ’ or ‘ \cap ’, $\phi_{B_{X_1}^R \circ B_{X_2}^R}(x) = |[x]_R|$.

Proposition 5. If for any non-empty set X , $B_{X_1}^R$ and $B_{X_2}^R$ be any two M-type bag structures drawn from X , then the following holds:

- $S(B_{X_1}^R \cup B_{X_2}^R) = S(B_{X_1}^R) \cup (B_{X_2}^R)$
- $S(B_{X_1}^R \cap B_{X_2}^R) = S(B_{X_1}^R) \cap (B_{X_2}^R)$

Proposition 6. If for any non-empty set X , $B_{Y_1}^R$ and $B_{Y_2}^R$ be any two M-type bag structures drawn from X , then for any $x \in (B_{Y_1}^R \cup B_{Y_2}^R)$, $\phi_{B_{Y_1}^R}(x) = \phi_{B_{Y_2}^R}(x)$.

4 Conclusion

This paper discusses the notion of bags and introduces a structure called M-type bag structure that is defined on a non-empty set associated with an indiscernibility relation. The characteristics of M-type bag structures together with their possible application are discussed and some of their algebraic properties are studied.

Acknowledgement. The author is grateful to the referees for their valuable comments which helped the preparation of the final version of this paper.

References

1. Blizard, W.D.: Multiset Theory. Notre Dame Journal of Formal Logic 30, 36–66 (1989)
2. Biswas, R.: An Application of Yager’s Bag Theory in Multicriteria Based Decision Making Problems. International Journal of Intelligent Systems 14, 1231–1238 (1999)

3. Chakrabarty, K.: A Note on the Effect of Knowledge Refinement on Bag Structures. In: Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS, vol. 6401, pp. 281–287. Springer, Heidelberg (2010)
4. Chakrabarty, K., Despi, I.: n^k -bags. *International Journal of Intelligent Systems* 22(2), 223–236 (2007)
5. Chakrabarty, K., Biswas, R., Nanda, S.: On Yager's theory of Bags and Fuzzy bags. *Computers and Artificial Intelligence* 18(1), 1–17 (1999)
6. Chakrabarty, K.: Notion of Fuzzy IC-Bags. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 12(3), 327–345 (2004)
7. Chakrabarty, K.: Decision Analysis using IC-Bags. *International Journal of Information Technology and Decision Making* 3(1), 101–108 (2004)
8. Chakrabarty, K.: The Notion of Intuitionistic Fuzzy IC-Bags. *Foundations of Computing and Decision Sciences* 37(1), 25–38 (2012)
9. Chakrabarty, K.: Bags with Interval Counts. *Foundations of Computing and Decision Sciences* 25(1), 23–36 (2000)
10. Chakrabarty, K.: On Bags and Fuzzy Bags. In: John, R., Birkenhead, R. (eds.) *Advances in Soft Computing, Soft Computing Techniques and Applications*, pp. 201–212. Physica-Verlag, Heidelberg (2000)
11. Delgado, M., Martin-Bautista, M.J., Sanchez, D., Vila, M.A.: On a Characterization of Fuzzy Bags. In: De Baets, B., Kaynak, O., Bilgiç, T. (eds.) IFSA 2003. LNCS, vol. 2715, pp. 119–126. Springer, Heidelberg (2003)
12. Delgado, M., Martin-Bautista, M.J., Sanchez, D., Vila, M.A.: An Extended Characterization of Fuzzy Bags. *International Journal of Intelligent Systems* 24, 706–721 (2009)
13. Rocacher, D., Bosc, P.: The Set of Fuzzy Relative Integers and Fuzzy Bags. *International Journal of Intelligent Systems* 24, 677–696 (2009)
14. Yager, R.R.: On the Theory of Bags. *International Journal of General Systems* 13, 23–37 (1986)

Coupled Tissue Bleeding Simulation in Virtual Surgery

Cheng Yang, Jiaxiang Guo, Jie Han, Xiangyun Liao, and Zhiyong Yuan*

School of Computer, Wuhan University, Wuhan, Hubei 430072
{ironyoung2012,whu18707199852,xyunliao}@gmail.com,
1024974785@qq.com, zhiyongyuan@whu.edu.cn

Abstract. Tissue bleeding simulation is the key problem in virtual surgery, which greatly reduces the risk of real surgery. It also plays an important role in the application of fluid-solid coupling. Traditional methods of fluid-solid coupling are almost mesh-based, in this paper we proposed a tissue bleeding model as well as an improved collision detection and response algorithm based on SPH (Smoothed Particle Hydrodynamics). Our tissue bleeding model describes Navier-Stokes equations with SPH method and our tissue model is set as rigid body. Experimental results show that our coupled tissue bleeding simulation model possesses high fidelity and strong robustness.

Keywords: Tissue Bleeding Simulation, Fluid-Solid Coupling, Collision Detection and Response, SPH.

1 Introduction

Virtual surgery uses a variety of medical data and virtual reality technology to create a virtual environment for planning, training, and guiding in real surgical procedure. So virtual surgery and related research has become an international field with rapid development. In virtual surgery training systems, the tissue bleeding simulation could greatly improve the visual feedback. And in aneurysm surgery, fluid simulation can provide useful information on aneurysm hemodynamics and pathology [2-3]. Thus, tissue bleeding simulation can improve the quality of virtual surgery and reduce the risk of surgery. It is urgent to address the problem in solid and fluid modeling, and the coupling between them in virtual surgery [1].

The tissue bleeding simulation belongs to the field of physical simulation, where mesh-based methods are widely used. However, mesh-based methods are time-consuming and not appropriate to deal with free surfaces, movements on the boundary, large deformation and collision propagation. But meshless methods can effectively solve these problems [4]. SPH has been the most widely used meshless method in recent years. Its particle properties make it suitable for deformable objects simulation [5]. The application of SPH method has been very mature for fluid simulation [6], but for solid simulation it need to be improved [4] [7].

* Corresponding author.

After fluid and solid modeling, we need to handle fluid-solid coupling. In [8], fluid and solid particles are represented as a whole to find neighboring particles, but this approach results in unrealistic situation that fluid particles penetrates into solid. [9] sets ghost particles to handle coupling, but the setting procedure is quite complicated. [10] proposes an algorithm for collision detection between spheres and triangle planes. [11] modifies the normal and tangential components of particle's velocity to response particle collision. [12-13] proposed friction between fluid and solid boundary particles by employing the laminar artificial viscosity model. On the bases of [10-13], we present a novel tissue bleeding model and an improved collision detection and response algorithm to handle fluid-solid coupling, and it successfully avoid problems in previous works.

2 SPH-Based Tissue Bleeding Model

Our fluid model is based on Navier-Stokes equations and described by SPH method [5] [13]. The basic idea of SPH method is to replace continuum with a set of discrete particles. The interaction between particles results in complex object movements. The integral approximate expression of SPH method:

$$f(x) \approx \sum_j^N f(x_j) W(x-x_j, h) \frac{m_j}{\rho_j} \quad (1)$$

Where $f(x)$ is the relative attribution at particles, such as density, pressure forces and viscosity forces [13]; h is the support domain of the smooth kernel function; j is the particle in the support domain of current particle; m_j is the mass of particle j ; ρ_j is the density of particle j ; $W(x-x', h)$ is the smooth kernel function, which can be thought as a weight distribution function.

Then, we represent solid model as a stationary and non-penetrating rigid body. Normal vectors of model surface particles are already known.

3 Coupling Process

Our coupling process includes two parts: collision detection and collision response. We extended Kasper's work [10] and proposed an algorithm for collision detection. Because collisions are detected in each time step, many solid particles may collide with single fluid particle. So we need to choose the nearest one for collision response. Then we use friction to modify the normal and tangential components of particle's velocity for collision response. Complete pseudocode is provided in Table 1.

Table 1. Collision Detection and Response Algorithm

Algorithm 1: Collision Detection and Response

```

1. while simulating do
2.   for each fluid-particle  $i$  do
3.     find solid particles
4.   for each solid particle do
5.     check if fluid particle will collide with itself in  $\Delta t$ 
6.   end for
7.     if find nearest collision solid particle
8.       modify the normal and tangential velocity of  $v_i$ 
9.     end if
10.    update fluid particle position  $x_i$ 
11.  end for
12. end while

```

Our collision detection outcomes in current time step will engender effect in next time step, so fluid and solid particles may penetrate into each other. To avoid this unrealistic scene, we introduce ‘‘Swept Sphere’’, a virtual sphere which the center is particle, and the radius is the product of the particle’s velocity and one time step.

The key point of our collision detection algorithm is the detection between single fluid particle and single solid particle. It needs following two steps.

3.1 Collision Detection between Particles and Tangent Planes

The current position of fluid particle p_{fluid} is x_{cur} , and the next position is x_{next} . We obtain:

$$x_{next} = x_{cur} + \Delta t \cdot v_{cur} \quad (2)$$

Where Δt is time step, v_{cur} is current velocity of p_{fluid} .

We can use equation (2) to obtain particle position in next time step. So the problem is translated to whether the fluid particle ‘‘Swept Sphere’’ will collide with tangent planes of solid surface particles. To solve this problem, we define the distance between fluid particles and tangent planes of solid surface particles as signed distance d_{signed} . To calculate d_{signed} , we need the normal vector N of the solid particle p_{solid} , the position of fluid particle x_{fluid} and plane constant C :

$$d_{signed} = N \cdot x_{fluid} + C \quad (3)$$

The signed distance is positive when p_{fluid} is located in the front of tangent planes or exactly on tangent planes; otherwise, the signed distance is negative. For better understanding, we explain the calculation process in Figure 1:

In Fig.1, O is the origin of coordinate system, the length of OI is the absolute value of $N \cdot x_{fluid}$ and NI is the signed distance between p_{fluid} and the solid particles' tangent plane. The length of ON is plane constant, and it can be calculated as follow:

$$C = |ON| = -(x_{solid} - O) \cdot N \tag{4}$$

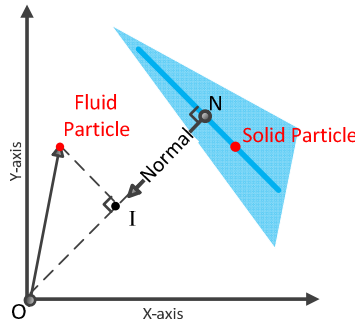


Fig. 1. The Schematic Diagram of Solving the Signed Distance (Blue: tangent planes of solid particles)

Then we calculate the time when fluid particle “Swept Sphere” intersecting with tangent planes. According to equation (3) (4), we can solve equation (5) by calculating as follows:

$$d_{signed}(x_{next}) = R \Rightarrow N \cdot (x_{cur} + t_0 * v_{cur}) + C = R \Rightarrow$$

$$t_0 * (N \cdot v_{cur}) + d_{signed}(x_{cur}) = R \Rightarrow t_0 = \frac{R - d_{signed}(x_{cur})}{N \cdot v_{cur}} \tag{5}$$

Where R is the radius of fluid particle “Swept Sphere”, and t_0 is the first time “Swept Sphere” collided with tangent planes. Similarly, t_1 is the last time the fluid particle collided with tangent planes:

$$t_1 = \frac{-R - d_{signed}(x_{cur})}{N \cdot v_{cur}} \tag{6}$$

If t_0 and t_1 are in the same time step, we decide that p_{fluid} will collide with solid particles' tangent planes in next time step, then we need further collision detection. Otherwise, it can be concluded that no collision between the fluid and solid particles.

3.2 Further Collision Detection between Particles

When it is determined that p_{fluid} would collide with a tangent plane, we can't directly determine whether the collision between the fluid and solid particle would happen

because the tangent planes can be extended infinitely. We need to further detect whether p_{solid} is inside “Swept Sphere”:

$$|x_{solid} - x_{fluid}| \leq R \tag{7}$$

If equation (7) is satisfied, we can conclude that p_{fluid} would collide with p_{solid} in next time step. Otherwise, they would not.

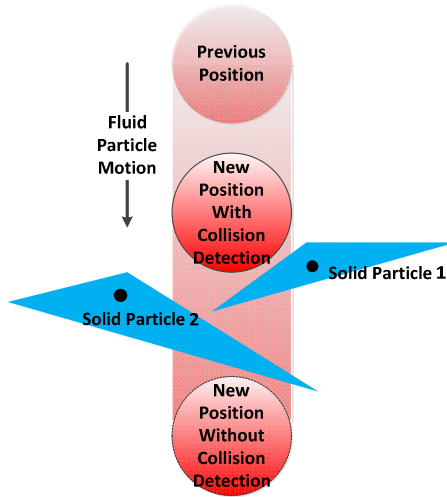


Fig. 2. The fluid particle may collide with Solid Particles 1 and 2, actually only with Particle 1 (Red: Fluid particle in “Swept Sphere”, Blue: tangent planes of solid particles)

For a fluid particle, there may be more than one solid particle would collide with within a time step. But single fluid particle will only collide with single solid particle (Fig.2), we need to select the nearest one among detected solid particles to make the collision response.

3.3 Collision Response

Along the normal vector of detected solid particles, we decompose fluid particle’s velocity into normal velocity and tangential velocity [10]. For no-slip collision, the direction of normal velocity will reverse while the direction of tangential velocity unchanged. Because tissues can’t be penetrated, we assume that the normal velocity of fluid particle will be 0 after collision and we just handle tangential velocity v_T .

Fluid-Solid Friction Force

In blood coupling, the friction plays an important role. We generate friction between interacting fluid and solid particles by employing the laminar viscosity model used in [12-13]. It is written as

$$f_{fluid \leftarrow solid} = -K m_{solid} m_{fluid} \nabla W(x_{fluid} - x_{solid}, h) \tag{12}$$

Where K is the coefficient of friction between the fluid particles i and solid particles j , is given as

$$K = -\frac{\sigma h c_s}{2\rho_{fluid}} \left(\frac{\min(\Delta \mathbf{v} \cdot \Delta \mathbf{x}, 0)}{|\Delta \mathbf{x}|^2 + \epsilon h^2} \right) \tag{13}$$

Where σ is the viscosity coefficient between fluid and solid, c_s is the speed of the numerical propagation, $\Delta \mathbf{v}_{ij} = \mathbf{v}_{fluid} - \mathbf{v}_{solid} = \mathbf{v}_T$, $\Delta \mathbf{x} = \mathbf{x}_{fluid} - \mathbf{x}_{solid}$, and ϵ is used to avoid singularities for $|x_{ij}| = 0$.

Velocity Updating

After calculating friction forces, we can compute the speed v_{next} of fluid particles in next time step by the momentum conservation in tangential direction:

$$m_{fluid} v_{next} = m_{fluid} \mathbf{v}_{T-cur} - f_{fluid \leftarrow solid} \cdot \Delta t \tag{14}$$

4 Experimental Results

Our experimental platform includes: (1) hardware: Intel (R) Core i5M460 2.53GHz CPU, 2GB memory, NVIDIA GeForce 320 GT display card; (2) software: Visual Studio 2010, OpenGL 4.2. We use an improved marching cubes algorithm [14] to triangulate the iso-surface and visualize the free surface to avoid the unresolved ambiguity. We store the neighboring eight voxels of each voxel in a linked list for the neighbors’ retrieval rather than global search to save time.

We simulate a scene of hand bleeding, blood is represented by red particles and the hand particles are yellow. Our hand model is constituted by 6986 particles, the time step $\Delta t = 0.0005s$ in our simulation clock, and the bleeding rate is $1p/5\Delta t$. Blood blows out from the injured hands, then flows slowly to the fingertips with the spread of the entire back of the hand. We both present the particle-based and rendered results in Fig. 3 and Fig.4. By the experimental data available, our simulation rate is averaged $368.2ms$ per step in particle-based simulation and is averaged $27.1ms$ per step in rendered simulation.

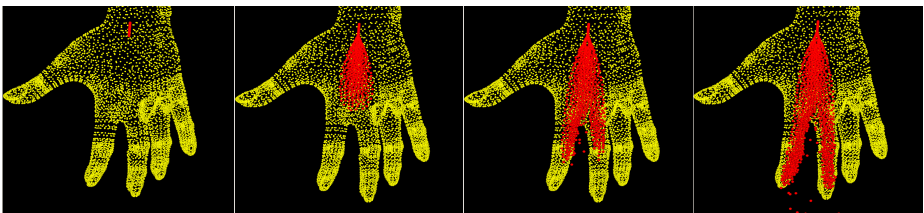


Fig. 3. Particle-based Hand Bleeding Simulation

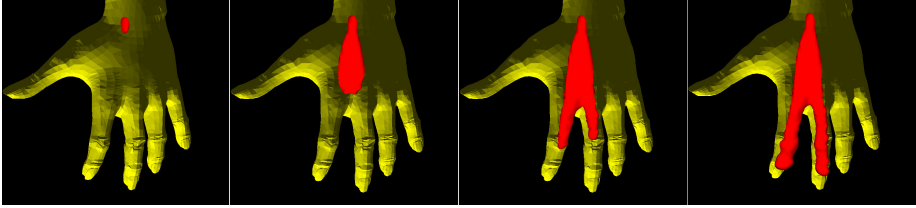


Fig. 4. Rendered Hand Bleeding Simulation

Comparison to [8]

Coupling algorithm in [8] has reasonable effects in soft fluid-solid collision, but it will result in unrealistic situation that fluid particles penetrate into solid particles in intense collision. Our coupling algorithm effectively avoids occurrence of this simulation.

Comparison to [9]

Our coupling algorithm only needs rigid body boundary layer of particles, but Ghost SPH needs all particles in rigid body to set Ghost Particle. And we just need the nearest solid particle for coupling simulation without complicated setting process of Ghost Particle in order to get less time consuming and smaller complexity.

Comparison to [10]

In collision response process between blood and tissue, we introduce the blood friction and momentum conservation equation to update the blood particle velocity. Compared to the method that normal and tangential velocity multiplied by a scaling factor [10], our method is accordance with the laws of momentum conservation.

5 Conclusion and Future Work

In this paper, we present a novel SPH-based tissue bleeding model for tissue bleeding simulation of high fidelity and great robustness. We propose a novel collision detection and response algorithm applies to this model, and our coupling algorithm has more advantages than the previous method. However, our collision detection algorithm will be invalid when the distribution of the surface particles on solid models is too sparse. And our simulation is not real-time enough. In the future, we will continue in-depth study of fluid-solid coupling model with sparse solid model, moving solid model, as well as the deformable tissue bleeding coupling. And we will introduce high-performance computing to our SPH-based tissue bleeding model to accelerate our simulation.

Acknowledgment. This work was supported by a grant from the National Natural Science Foundation of China (Grant No. 61272276).

References

1. Xing, Y.J., Zhang, S.H., Liu, X.B.: The Present Situation of Virtual Surgery System Technology. *Computer Engineering and Applications* 7, 88–145 (2004)
2. Müller, M., Schirm, S., Teschner, M.: Interactive Blood Simulation for Virtual Surgery Based on Smoothed Particle Hydrodynamics. *Technology and Health Care* 12(1), 25–31 (2004)
3. Dwyer, H.A., Cheer, A.Y., Barakat, A.I., et al.: Unsteady And Three-Dimensional Simulation of Blood Flow in the Human Aortic Arch. *Transactions of the ASME* 124, 378–387 (2002)
4. Müller, M., Keiser, R., Nealen, A., et al.: Point Based Animation of Elastic, Plastic And Melting Objects. In: *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 141–151. Eurographics Association (2004)
5. Liu, G.R., Liu, M.B.: *Smoothed Particle Hydrodynamics—A Meshfree Particle Method*. World Scientific Publishing, Singapore (2003)
6. Lucas, B., Thomas, L.: *An initiation to SPH*. Department of Mathematics, PUC–Rio – Rio de Janeiro, Brazil (2009)
7. Keiser, R., Adams, B., Gasser, D., et al.: A Unified Lagrangian Approach to Solid-Fluid Animation. *Point-Based Graphics*. In: *Eurographics/IEEE VGTC Symposium Proceedings*, pp. 125–148. IEEE (2005)
8. Barbara, S., Jurg, S., Renato, P.: A Unified Particle Model for Fluid-Solid Interactions. *Computer Animation and Virtual Worlds* 18(1), 69–82 (2007)
9. Hagit, S., Robert, B.: Ghost SPH for Animating Water. *ACM Transactions on Graphics (TOG)* 31(4), 1–8 (2012)
10. Fauerby, K.: *Improved Collision detection and Response* (2003), <http://www.peroxide.dk>
11. Afonso, P., Fabiano, P., Thomas, L., Geovan, T.: Particle-based Viscoplastic Fluid /Solid Simulation. *Computer-Aided Design* 41(4), 306–314 (2009)
12. Becker, M., Teschner, M.: Weakly compressible SPH for free surface flows. In: *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 209–217. Eurographics Association (2007)
13. Monaghan, J.J.: *Smoothed Particle Hydrodynamics*. *Reports on Progress in Physics* 68(8), 1703 (2005)
14. Jin, J., Wang, Q., Shen, Y., Hao, J.S.: An Improved Marching Cubes Method for Surface Reconstruction of Volume Data, *Intelligent Control and Automation*. In: *WCICA 2006, The Sixth World Congress*, vol. 2, pp. 10454–10457 (2006)

The Study of Adaptive Modulation Algorithms in OFDM System

Yanli Shi^{1,*}, Yanbin Shi¹, Renxia Ou², and Haixia Yu¹

¹ Information Countermeasure Department, Air Force Aviation University,
Changchun, Jilin, China 130022
shiyianli0607@163.com

² Public Health Department, Jilin Medical College, Jilin, Jilin, China 132013
yusong95599@sohu.com

Abstract. The adaptive modulation techniques is widely used in orthogonal frequency division multiplexing (OFDM) system, this technique can not only improve the utilization efficiency of spectrum, but also the bit error rate performance of system can be significantly improved. Adaptive modulation algorithm is important part in achieving performance improvement, this paper introduces three adaptive bit and power modulation algorithms based on the different allocation principle, and through simulation the paper can comparatively analyze the performance of different algorithms.

Keywords: OFDM, adaptive performance, modulation, algorithms.

1 Introduction

Frequency band is the most valuable resources in wireless communication, along with people has rapid demand to the wireless data services, but how to realize the maximal data transmission rate in the limited bandwidth, that is to say how to realize maximal utilization efficiency of frequency band, it becomes the research focus of mobile communication. Due to adaptive technology has strong advantage in increasing data transmission rate and utilization rate of spectrum, thus it becomes the key technology in the current and future mobile communication system.

2 OFDM Principle

The OFDM technology is different from the FDM technology, its basic principle: the high speed data stream pass the transform of tandem and parallel connection, and the data stream be assigned to the several sub-channels that its transmission rate is relatively low; last the data stream can transmit in the channel. Because of the symbol period of each channel increase relatively, so it can lessen the impact that the multi-path time delay spread of the wireless channel have affection on the system. and also the protection interval is inserted between the OFDM symbols, and let the protection interval is greater than the maximum delay spread of wireless channel, thus can

maximally eliminate ISI produced by the multi-path interference, generally using cyclic prefix as guard interval, which can be avoided the ICI produced by multi-path channel interference. The basic block diagram of OFDM system is shown in Fig.1.[1-3]

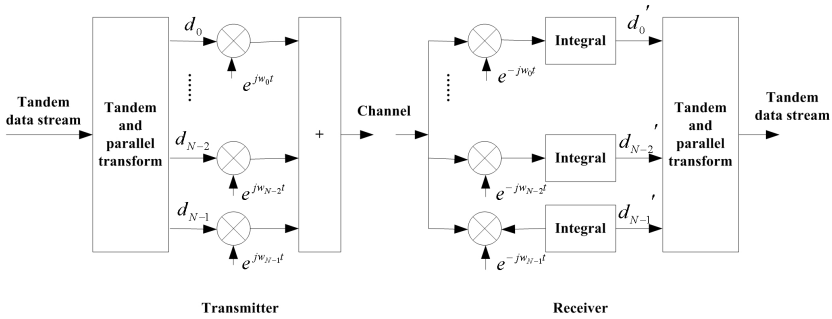


Fig. 1. Block diagram of the basic principles of OFDM system

From the diagram we can see: in the transmitter of OFDM system, the basic principle is that the input data transform N sub-channels data through tandem and parallel connection transform, then each orthogonal sub-carrier respectively modulated correspondingly and stacking synthetic output together. In the receiver of OFDM system, after each sub-carrier respectively mixing and integrating we will get various data, and output the original data through serial conversion. The fundamental idea of the OFDM technology is that the serial high speed data stream change into the parallel low speed data streams through the tandem and parallel transform, the key point is to ensure the orthogonal performance of each sub-carrier.

3 Adaptive OFDM Principle

The equivalent base-band of adaptive OFDM system model is shown in Fig.2. Its basic principle is summarized as follows: first, the receiving end obtains the channel state information through the channel estimator; according to the built-in algorithm and channel state information from the channel estimator, the adaptive bit and power divider sets appropriate modulation parameters for each sub-carrier, (including two parts of the modulation mode and sending power), and these parameters are putted into the transmitting end through dedicated channel; according to the modulation parameters, the tandem-parallel converter of sending end allocates corresponding bits for every sub-carrier, then the modulator completes base-band modulation of each sub-carrier; the output data of each modulator added into the channel and transmitted to receiving end through the fast IFFT converter, parallel / tandem converter, adding cyclic prefix, the receiving end process the inverse operation comparing to the sending end, then the final data is outputted.[4-7]

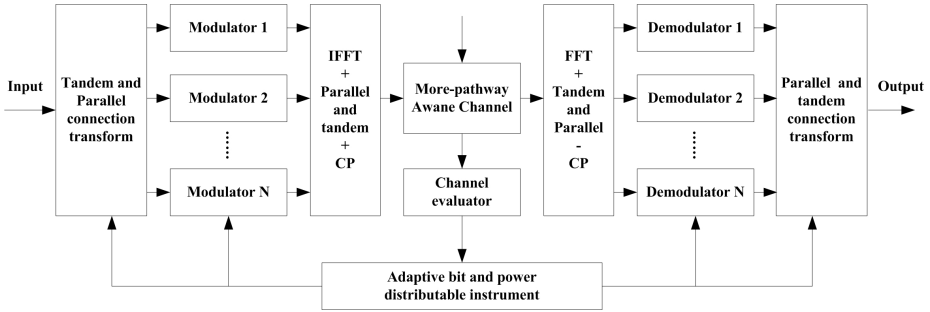


Fig. 2. Base-band model diagram of adaptive OFDM system

From the above principles, in order to the whole system can obtain good performance, the key parts are the channel estimator, adaptive bit and power distributor, when the channel state information of the channel estimator is more accurate, the performance of adaptive bit and power divider is more better.

4 Adaptive Modulation Algorithms

The performance of adaptive OFDM system, the key two parts are the channel estimator, adaptive bit and power distributor, and the adaptive algorithms of two parts are the core parts, this paper presents three different adaptive bit and power optimum modulation algorithms based on three different allocation principles: the Hughes-Hartogs algorithm based on minimum power, the Chow algorithm based on maximal system capacity and the Fischer algorithm based on minimal error rate. And this paper also presents improvement that the complexity of Hughes-Hartogs algorithm and the performance of Fischer algorithm.[8-13]

4.1 Hughes-Hartogs Algorithm

The Hughes-Hartogs algorithm is an adaptive bit and power allocation algorithm that applied to the OFDM system or multi-carrier system, it first used in downlink high speed data transmission of xDSL cable system. This algorithm is based on the channel gain, its basic idea: the bit initial value sets to 0 in each sub channel, and then all unallocated bits are allocated to the corresponding sub-channel. Each allocation, first the algorithm find the increase of 1 bit, and the minimal sub-channel needing to increase power, then the bit number of sub-channel is increased 1. So cycling, until all of the bits are dispensed, the required power of each sub-channel final calculated.

The Hughes-Hartogs algorithm can achieve the optimal bit and power allocation results, but require additional search and sorting, the computational complexity is great. Especially when the sub channel number is more, the Hughes-Hartogs algorithm is very slow.

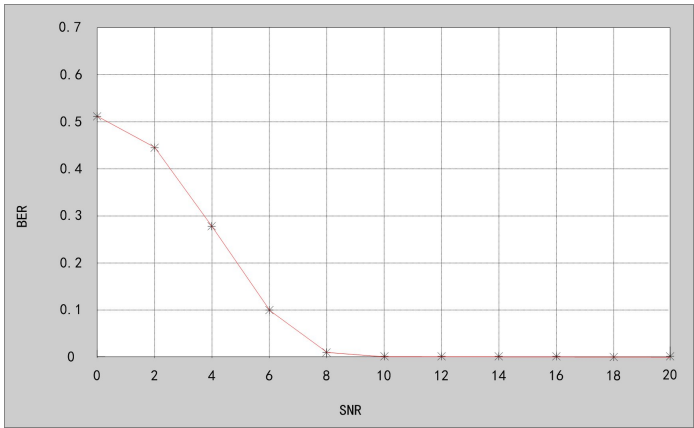


Fig. 3. BER curves of Hughes-Hartogs algorithm

Fig.3 shows that the Hughes-Hartogs algorithm based on minimal power distribution maintains good system performance, it can be said to be a classic algorithms, but its complexity is very high, the calculation amounts is large, because calculation time is too long so that the algorithm cannot get good application in the actual system. Due to the Hughes-Hartogs algorithm has such shortcomings, the algorithm can be improved, the improved Hughes-Hartogs algorithm can reduce allocation complexity within the premise that the performance maintains in the allowable range, and it can overcome shortcomings.

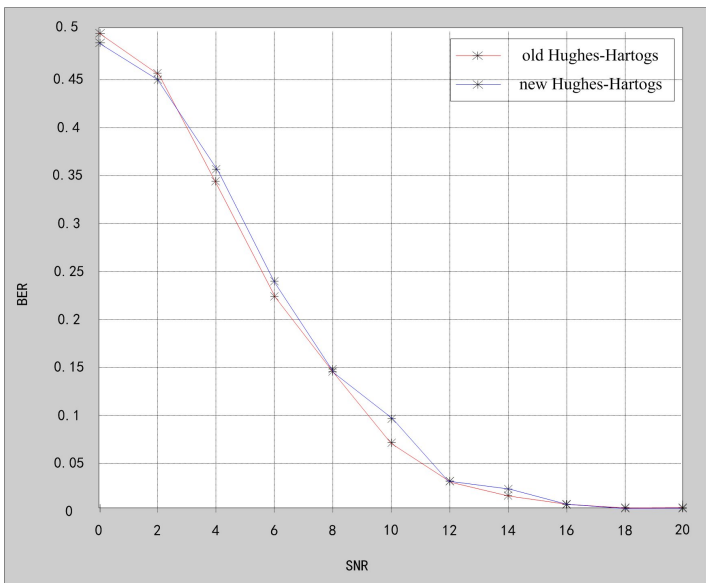


Fig. 4. BER curves of improved Hughes-Hartogs algorithm

The system BER simulation curve as shown in Fig.4, the BER curves shows that: the improved adaptive modulation bit allocation algorithm has not very big influence on the bit error rate (BER) performance of the system, when the SNR value is same; the improved adaptive algorithm reduces about one order of magnitude comparing to the original algorithm.

4.2 Chow Algorithm

Comparing to the Hughes-Hartogs allocation algorithm, the Chow algorithm greatly reduces the complexity of the algorithm, improves the speed of the bit allocation, and the adaptive modulation can be used in high speed data transmission. The Chow algorithm allocates bit based on the channel capacity of sub-channel. Its optimization criterion makes that the system spectrum efficiency achieves best under the condition maintaining the target bit error rate (BER). The algorithm uses an iterative process, gradually allocates bit, also makes the system margin gradually increased, until the entire bit allocated.

4.3 Fischer Algorithm

The Fischer algorithm is an improved algorithm based on the Chow algorithm. Comparing to the Chow algorithm, the basis of the Fischer algorithm bit allocation is not the channel capacity, its design goal is the bit error rate (BER) minimum. Its optimization criterion is that the bit error rate (BER) of the system achieves optimal performance under the premise of maintaining the constant transmission rate and the total transmission power.

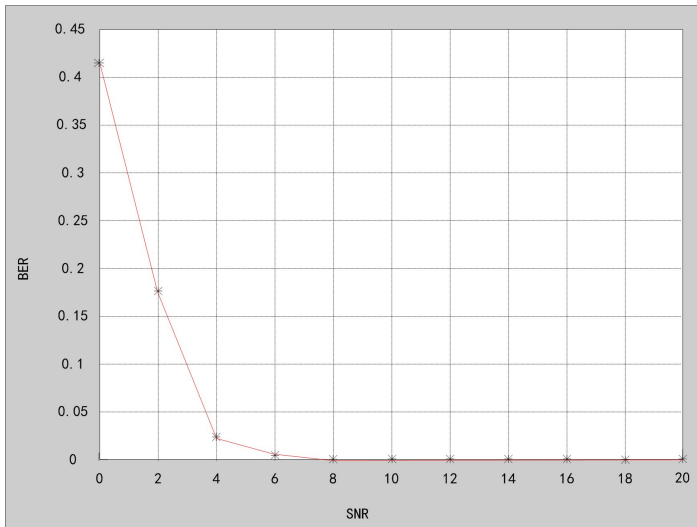


Fig. 5. BER curves of Fischer algorithm

Fig.5 shows that this OFDM adaptive modulation algorithm based on the minimum bit error rate (BER) principle can meet the requirements in system performance, from the comparison of figure 3 and Fig.5, the BER performance of the Fischer algorithm is better than the Hughes-Hartogs algorithm based on the minimum power allocation principle. So the Fischer algorithm is applied to the businesses that bit error rate (BER) demanding is high, such as data transmission, multimedia services and so on, and it has been adapted in xDSL system. This algorithm also has some shortcomings, the shortcomings can be improved, the improved Fischer power allocation algorithm directly distribute power according to the quantized scale bits, so as it can reduce the quantization error of the power allocation, and get better performance of allocation affection than the original algorithm.

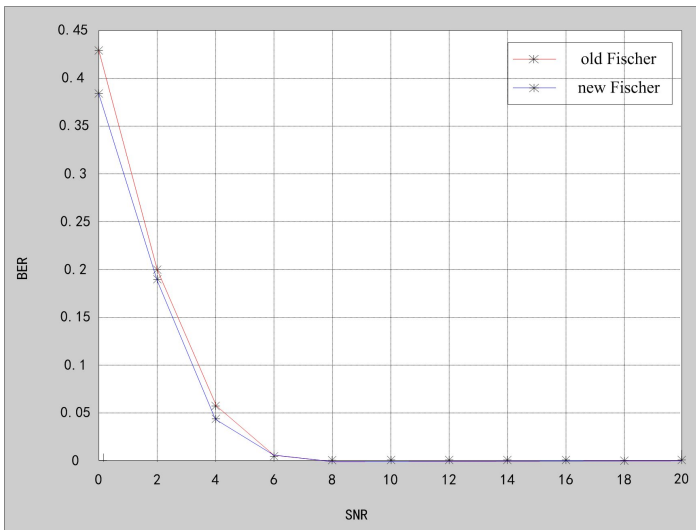


Fig. 6. BER curves of improved Fischer algorithms

Fig.6 shows, when the SNR value is same, the error rate of this improved power allocation algorithm is reduced somewhat than the original algorithm, in low SNR environment the error rate probably drop about 0.03.

5 Conclusions

This paper introduces three adaptive modulation algorithms based on different optimal criterion, then the Fischer algorithm and the Hughes-Hartogs algorithm also is improved, from the simulation results we can see the respective improvement results of different algorithms, the complexity of the improved Hughes-Hartogs algorithm is better than the original algorithm, but the performance is reduced slightly, so that it is possible in the actual application. The performance of the improved Fischer algorithm is improved than the original algorithm, and the system bit error rate is further reduced.

References

1. Song, Z.F., Zhang, K.L., Guan, Y.L.: Statistical adaptive modulation for QAM-OFDM systems. *IEEE GLOBECOM*, 706–710 (2002)
2. Shi, Y., Li, Y., Wang, J.: Effectiveness evaluation about self-screening jamming of assaulting aircraft in complex EM environment. In: Yu, Y., Yu, Z., Zhao, J. (eds.) *CSEEE 2011, Part II. CCIS*, vol. 159, pp. 360–365. Springer, Heidelberg (2011)
3. Li, Z., Wang, W.H., Zhou, W.N., Song, J.D.: A Simple Bit and Power Allocation Algorithm for OFDM Systems. In: *Proceedings of ICCT*, pp. 60–70 (2003)
4. Dardari, D.: Ordered sub-carrier selection algorithm for OFDM-based high-speed WLANs. *IEEE Trans.* 3, 1452–1458 (2004)
5. Song, K.B., Ekbal, A., Cioffi, J.M.: Adaptive modulation and coding (AMC) for bit-interleaved coded OFDM (BIC-OFDM). In: *Proc. IEEE ICC 2004-I*, vol. 6, pp. 3197–3201 (2004)
6. Benvenuto, N., Tosato, F.: On the selection of adaptive modulation and coding modes over OFDM. In: *Proc. IEEE ICC 2004*, vol. 6, pp. 3251–3255 (2004)
7. Hwang, Y.T., Tsai, C.Y.: T, L., Cheng, C: Block-wise adaptive modulation for OFDM WLAN systems. In: *Proc. IEEE ISCAS 2005*, vol. 6, pp. 6098–6101 (2005)
8. Guocong, S., Ye, I.: Cross-layer optimization for OFDM wireless networks-part I: theoretical framework. *IEEE Trans. Wireless Commune.* 4, 614–624 (2005)
9. Zhou, K.N., Yong, H.C.: Heuristic Algorithms to Adaptive Sub-carrier and Bit Allocation in Multi-class Multi-user OFDM Systems. In: *Vehicular Technology Conference 2006*, vol. 3, pp. 1416–1420 (2006)
10. Phamdin, T., Lee, W., Cottonet, M., Jitapunkul, S.: Novel V-BLAST-based Adaptive Modulation for MIMO-OFDM Systems over Frequency Selective Fading Channels Information. In: *Communications and Signal Processing 2005 Fifth International Conference*, pp. 1227–1230 (2005)
11. Xia, X.T., Wang, Z.G.: Expanded uncertainty model based on whitened weighting function. In: *Proceedings of the Second International Symposium on Instrumentation Science and Technology*, vol. 3, pp. 935–938 (2002)
12. Fasano, A.: On the optimal discrete bit loading for multi-carrier systems with constraints. In: *Proc. IEEE VTC 2003*, pp. 915–919 (2003)
13. Wang, H.L.: Grey Cloud model and its application in intelligent decision support system supporting complex decision. In: *International Colloquium on Computing, Communication, Control, and Management*, vol. 1, pp. 542–546 (2008)

An Improved Algorithm of Parameters Estimation for Frequency-Hopping Signal

Jun Lv*, Weitao Sun, and Tong Li

Department of Information Engineering,
Academy of Armored Forces Engineering Beijing, China
344243278@qq.com

Abstract. In view of the problem of S-transform that the width of the window shortens quickly with the increase of frequency, and the performance of parameter estimation of FH (Frequency-Hopping) decline, an improved algorithm based on asymmetric window is proposed. The algorithm firstly studies the window function of S-transform, and selects hyperbola of window function. This window function suits for estimation of FH parameter. Because rising and falling edge of this window function has large degree of freedom, on the other hand the asymmetry of the window function varies with frequency. Considering the distortion of the ridge line at high frequency component, and combining with the statistical property of white Gaussian noise in S spectrum, an indirect band-pass filtering method in time-frequency plane is proposed to smooth the ridge line and decline the influence of noise. Theoretical analysis and simulation results verify the effectiveness and feasibility of the proposed algorithm.

Keywords: S-transform, Asymmetrical window function, Hopping signal, Parameter estimation.

1 Introduction

The frequency hopping (FH) signal is one of the Spread Spectrum Methods and widely used in modern military communications. For the strong anti-jamming, low probability of intercept and excellent networking capabilities, the FH signal is also a great challenge to communication countermeasure field. Thus, Interception, detection and blind parameter estimation of the FH signal are urgent problems in modern communication countermeasure field.

In 1996, the American geophysicists Stockwell et al propose S-transform, which is a method of time-frequency analysis between STFT (Short Time Fourier Transform) and wavelet transform [1]. Considering S-transform's perfect localizing characteristic in time-frequency domain, this paper propose an improved S-transform based on asymmetric window function. This paper gives the parameter estimating algorithms, obtains more exact estimating results by computer simulating, and the simulating results show that the proposed method could estimate the FH parameters effectively.

* Corresponding author.

2 Original S-Transformation

The definition of S-transform can be deduced by the STFT and wavelet transform. The S-transform of $x(t)$ is defined as [2]:

$$STFT_x(t, f) = \int_{-\infty}^{+\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau} d\tau \tag{1}$$

Where $w(t)$ is window function. The choice of the window function decides the resolution STFT. According to the Heisenberg Uncertainty Principle, the areas of Time-Frequency image can reach minimum when the window function is Gaussian window. Therefore, the window function of (1) is defined as Gaussian window:

$$w(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{t^2}{2\sigma^2}\right] \tag{2}$$

Where σ is scale factor of time width. Since the signal is non-stationary signals, so we hope the time resolution can be adjusted. At higher frequencies, the signal fiercely change, and the mutative period of the time is relatively small, so time window should be taken to narrow. At lower frequencies, the signal is relatively stable, and the mutative period of the time is relatively large, so time window should be taken to wider. Thus, the σ defined as:

$$\sigma(f) = \frac{1}{|f|} \tag{3}$$

So forming formula (1), (2) and (3), we can define S-transform as:

$$S_x(t, f) = \frac{|f|}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x(\tau)e^{-\frac{(\tau-t)^2 f^2}{2}} e^{-j2\pi f\tau} d\tau \tag{4}$$

3 The Research of Window Function of S-Transform

The definition of S-transform of (4) can be redefined as:

$$S_x(t, f) = \int_{-\infty}^{+\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau} d\tau \tag{5}$$

Where $x(t)$ is signal, and $w(t)$ is window function. In order to ensure the Inverse transformation of S-transform is the only, the window function required to satisfy the following conditions:

$$\int_{-\infty}^{+\infty} w(t)dt = 1 \tag{6}$$

By using the method of symmetrical windows, if it is necessary to improve the resolution of time, it can only be achieved by reducing the width of window. But it will make the resolution of frequency sharply decline. Conversely, if we want to increase

the resolution of frequency, the resolution of time will sharply decline. This will seriously affect the estimated effect of parameter. Therefore, CR Pinnegar et al introduced the concept of asymmetrical window [6][7][8].

3.1 The Hyperbola Window

In S-transform, if the width of window of time domain is same, the higher the resolution of frequency can be provided when the window function is more symmetrical. Conversely, the higher the resolution of time can be provided when the window function is more asymmetrical. Therefore, in the low frequency region, we should choice the asymmetrical windows to improve the resolution of time, and in the high frequency region, we should choice the symmetrical windows to improve the resolution of frequency. Thus, we can define the window function as:

$$w_{HY}(t-\tau) = \frac{|f|}{\sqrt{2\pi}} \frac{2}{r_{HY}^B + r_{HY}^F} \exp\left(\frac{-f^2 \left[C(t-\tau, \{r_{HY}^B, r_{HY}^F, \lambda_{HY}^2\}) \right]^2}{2}\right) \quad (7)$$

$$C(t-\tau, \{r_{HY}^B, r_{HY}^F, \lambda_{HY}^2\}) = \left(\frac{r_{HY}^B + r_{HY}^F}{2r_{HY}^B r_{HY}^F}\right)(t-\tau-\xi) + \left(\frac{r_{HY}^B - r_{HY}^F}{2r_{HY}^B r_{HY}^F}\right)\sqrt{(t-\tau-\xi)^2 + \lambda_{HY}^2}$$

$$\xi = \sqrt{\frac{(r_{HY}^B - r_{HY}^F)\lambda_{HY}^2}{4r_{HY}^B r_{HY}^F}} \quad (9)$$

Where, $C(t-\tau, \{r_{HY}^B, r_{HY}^F, \lambda_{HY}^2\})$ is hyperbola function, r_{HY}^F is the parameter of rising edge, r_{HY}^B is the parameter of falling edge ($0 < r_{HY}^F < r_{HY}^B$), and λ_{HY} is the parameter of positive curvature. In order to make the center of $w_{HY}(t-\tau)$ always kept at $t-\tau=0$, we introduce a parameter ξ . We make a simulation for the hyperbola window, he results are shown Fig.1

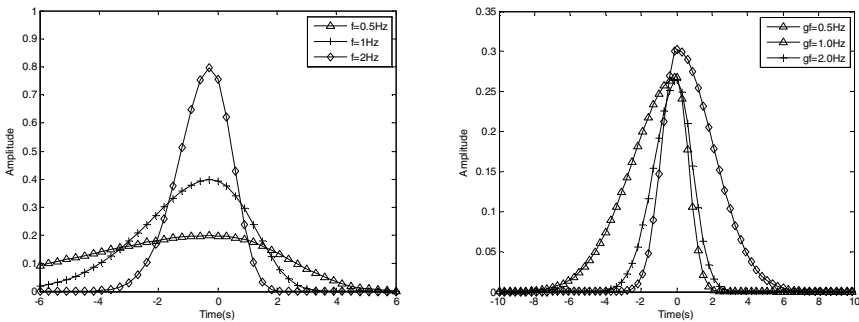


Fig. 1. Different roles of the various parameters for Hyperbola window

From the Fig 2, we can see the more the frequency is small the more the asymmetry of the window function is the strong. However, as the frequency increased, the asymmetry of the window function gradually weakens, and more and more closes to the Gaussian window. The effect of curvature parameter: λ_{HY} is bigger, the symmetry of window is changes more quickly with frequency changing. From the Fig 3, we can see the more r_{HY}^F (gf) is big, the more the rising edge is steep. As the same as the more r_{HY}^B (gb) is big, the more the falling edge is steep

3.2 Double Gaussian Window

Similar to the hyperbolic window, we can define the double Gaussian window as:

$$w_{BG}(t - \tau, f, \{r_{BG}, r_{BG}^F\}) = \frac{|f|}{\sqrt{2\pi}} \frac{2}{r_{HY}^B + r_{HY}^F} \exp\left(\frac{-f(t - \tau)^2}{2[\tilde{r}_{BG}(t - \tau)]^2}\right) \quad (10)$$

$$\begin{aligned} \tilde{r}_{BG}(t - \tau) &= r_{BG}^B, \quad \tau \geq t \\ &= r_{BG}^F, \quad \tau < t \end{aligned} \quad (11)$$

$$r_{BG}^B = \frac{\pi - 4}{2(\pi - 2)} r_{BG}^F + \frac{\sqrt{(8\pi - 3\pi^2)(r_{BG}^F)^2 + (4\pi^2 - 8\pi)r_{BG}}}{2(\pi - 2)} \quad (12)$$

Where r_{BG} , r_{BG}^F are constant. r_{BG} can adjust the double-width of Gaussian window. r_{BG} is more larger, the width of Gaussian window is more bigger. r_{BG}^F represents the degree of variation of the falling edge of the double Gauss window. r_{BG}^B can represent by r_{BG} and r_{BG}^F , like (16). We make a simulation for the double Gaussian window when the r_{BG}^F (gf) select different value, the results is shown Fig4

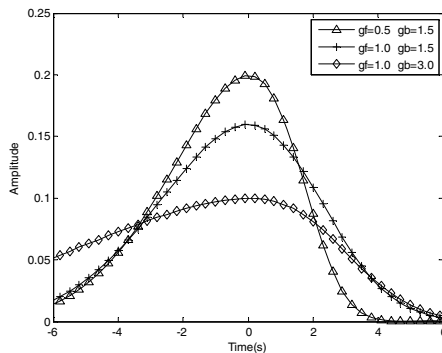


Fig. 2. Different roles of the various parameters for Double Gaussian Window

From the Fig4, we can see that with the increasing, the rising edge of the window function is more and more gentle, but the falling edge is more and more steep.

For the symmetric window, the rising edge changes too slowly to accurately estimate the frequency hop timing. And the asymmetrical window can effectively solve this problem. For estimation of parameter of the FH signal, accurately estimation of the hop timing is very important. Therefore, we should choose the asymmetrical window to estimate the parameter of the FH signal.

4 Parameter Estimation of FH Signal Using S-Transform with an Asymmetrical Window

The signal, after making s-transform, has better resolution of time at high frequency region. And in order to improve the resolution of frequency, we hope the wind function is more symmetrical. From the analyzing of the previous section, we know that the asymmetry of Hyperbola window is gradually weakened as the frequency increasing. But the asymmetry of the double Gaussian window doesn't change as the frequency increasing. Therefore, the hyperbola window is more suitable than double Gaussian window for estimation of parameters of FH.

The noise will affect peak sequences. Thus, in order to void it, this paper use this method, which extract time-frequency ridge, to estimate the parameters of FH signal. Algorithms are as follows:

Step 1: Setting the various parameters of the Hyperbola window function: $r_{HY}^F = 0.5$, $r_{HY}^B = 2$, $\lambda_{HY} = 2$. S-transform $x(t)$, getting $S_x(t, f)$.

Step 2: Extracting ridge in time-frequency plane, getting $l_r(t, f)$. And the $l_r(t, f)$ is actually the Time coordinate of the peak sequence. In MATLAB, The extraction method is as follows:

$$[l(t), l_r(t, f)] = \max[S_x(t, f)]$$

Where $l(t)$ are peak sequences.

Step 3: Smoothing of the $l_r(t, f)$. The $l_r(t, f)$ generated some larger peak pulse, since the of interference noise, so we need smooth the $l_r(t, f)$.

Step 4: After Smoothing of the $l_r(t, f)$, getting the derivative of $l_r(t, f)$ d along the time axis. We can get $l'_r(t, f)$, and set the threshold of ΔF . If the value of $l'_r(t, f)$ is greater than ΔF , the value is hop timing.

Step 5: According to the obtained hop timing, We can calculate the hop duration, which equal average value of the difference between the adjacent hop timings. The frequency of FH signal can be got by FFT the signal between the adjacent hop timings.

5 Simulation and Results Analysis

The parameters of the simulated FH signal:

- a. The hop duration: $T_h = 50ms$.
- b. The hop frequencies: {1.0, 1.2, 1.5, 0.9, 1.6, 1.4, 1.1, 1.3}.
- c. Hop timing of FH signal {50, 100, 150, 200, 250, 300, 350}.

The Fig 3(a) shows the original S-transform of simulated FH signal, and Fig 3(b) shows the improved S-transform of simulated FH signal.

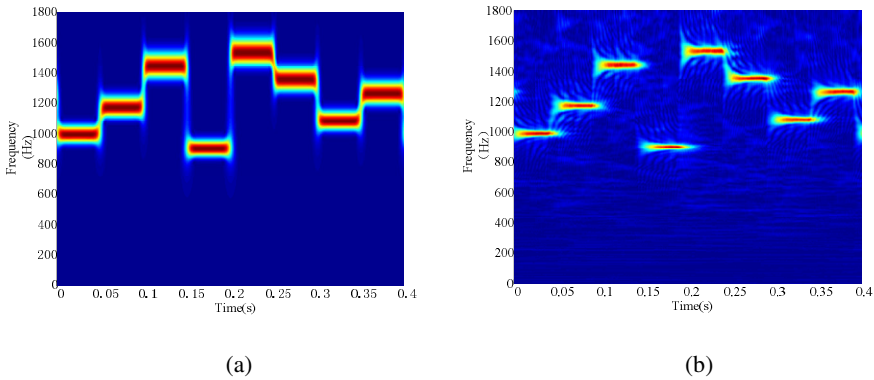


Fig. 3. Time-frequency distribution figure

From the Fig3, we can see that the aggregation of frequency-time and the resolution of time and frequency significantly enhance when the S transform use the asymmetric window function.

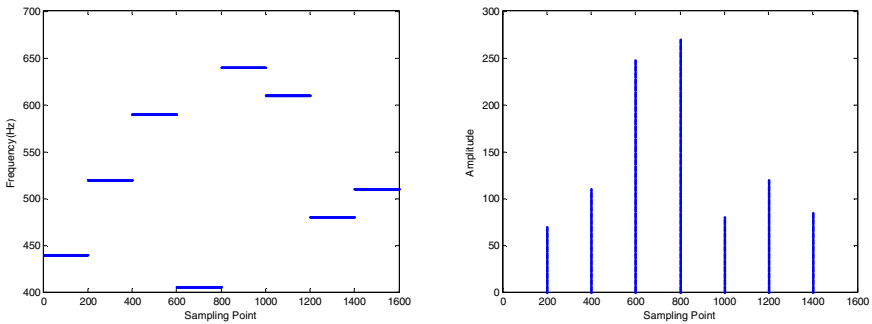


Fig. 4. The extraction of Time-frequency ridge and estimation for hop timing

The Fig4 shows smoothed the ridge of time-frequency, we can see the effect of smoothed is better. And the Fig 8 shows the hop timing by doing differential of the

ridge of time-frequency. From the Fig 8, we can see the simulated signals have 7 hop timing.

According to the algorithm, we calculate the parameters of FH Signal .and the results are as follows:

a. Hop timing αT

The Table 1 shows the estimated results of hop timings of 500 independent experiments, when the $SNR = 10\text{dB}$.From the Table 1, we can see the estimated results of hop timings are very closed to the real value. The average relative error is less than 0.1%.

b. Hop duration T_h

From the Table 1, we can calculate the period of FH signal. $T_h = 50.012\text{ms}$, and the average relative error is less than 0.023%

c. Hop Frequencies

The Table 2 shows the estimated results of the frequencies of FH signal of 500 independent experiments. From the Table 2, we can see the estimated results of frequencies are much closed to the real value. The average relative error is less than 0.4%.

Table 1. The estimated results of hop timings of 500 independent experiments Unit ms

	αT_1	αT_2	αT_3	αT_4	αT_5	αT_6	αT_7
Real Value	50	100	150	200	250	300	350
Average Value	50.08	99.83	149.95	200.11	250.19	299.84	350.15
Relative Error%	0.16	0.17	0.033	0.055	0.076	0.053	0.043

Table 2. The estimated results of the frequencies of FH signal of 500 independent experiments Unit KHz

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Real Value	1.0	1.2	1.5	0.9	1.6	1.4	1.1	1.3
Average Value	1.0032	1.2014	1.4954	0.9058	1.5921	1.4068	1.1073	1.2942
Relative Error%	0.320	0.117	0.307	0.644	0.494	0.486	0.664	0.446

6 Conclusions

In this paper, we study two types of window function of S-transform, which are symmetrical window and asymmetrical window. And in the asymmetrical window, we introduce two window functions, which are Hyperbola window and double Gaussian window. By comparison, we ultimately chose the Hyperbola window for the window function of S-transform. And the theoretical analysis and simulation results verify the effectiveness and feasibility of the proposed algorithm.

References

1. Stockwell, R.G.: S-Transform Analysis of Gravity Wave Activity from a Small Scale Network of Airglow Imagers. Physics Doctor Thesis. Physics Department, University of Western Ontario, London, On Canada (1999)
2. Pinnegar, C.R., Eatond, W.: Application of the S Transform to Prestack Noise Attention Filtering. *Journal of Geophysical Research* 108(B9), 2422–2431 (2003)
3. Sitanshu, S.S., Ganapati, P., Nithin, V.: An Improved S-Transform for Time-Frequency Analysis. In: *IEEE International Advance Computing Conference*, pp. 315–319 (2009)
4. Pinnegar, C.R., Mansinha, L.: The Bi-Gaussian S-transform. *Society for Industrial and Applied Mathematics* 24(5), 1678–1692 (2003)
5. Zhu, M.Z., Ji, H.B., Gao, R.T.: Parameter Estimation of Hybrid DS/FH Spread Spectrum Signals Using S Transform with an Asymmetrical Window. In: *International Conference on Communications and Mobile Computing*, pp. 329–322 (2010)
6. Nithin, V.G., Sitanshu, S.S., Mansinha, L.: Time Localised Band Filtering Using Modified S-Transform. In: *International Conference on Signal Processing Systems*, pp. 42–46 (2009)
7. Robert, P.C., Lalu, M.: The S-Transform with Windows of Arbitrary and Varying Shape. *Geophysics* 68(1), 381–385 (2003)
8. Brown, A.R., Lauzon, M.L.: A General Description of Linear Time-frequency Transforms and Formulation of a Fast, Invertible Transform That Samples the Continuous S-Transform Spectrum Non-redundantly. *IEEE Transactions on Signal Processing* 58(1), 281–290 (2010)

The Feature Extraction Method of EEG Signals Based on Degree Distribution of Complex Networks from Nonlinear Time Series

Fenglin Wang^{1,2}, Qingfang Meng^{1,2,*}, Weidong Zhou³, and Shanshan Chen^{1,2}

¹ School of Information Science and Engineering, University of Jinan, Jinan 250022, China

² Shandong Provincial Key laboratory of Network Based Intelligent Computing,
Jinan 250022, China

³ School of Information Science and Engineering, Shandong University, Jinan 250100, China
ise_mengqf@ujn.edu.cn

Abstract. The nonlinear time series analysis method based on complex networks theory gives a novel perspective to understand the dynamics of the nonlinear time series. Considering the electroencephalogram (EEG) signals showing different nonlinear dynamics under different brain states, this study proposes an epileptic EEG analysis approach based on statistical properties of complex networks and applies the approach to epileptic EEGs automatic detection. Firstly, the complex network is constructed from the epileptic EEG signals and the degree distribution (DDF) of the resulting networks is calculated. Then the entropy of the degree distribution (NDDE) is used as a feature to classify the ictal EEGs and the interictal EEGs. The experiment results show that the NDDE of the ictal EEG is lower than interictal EEG's and the classification accuracy, taking the NDDE as a classification feature, is up to 96.25%.

Keywords. EEG, Epilepsy, Seizure detection, Complex networks, Degree distribution, Shannon entropy.

1 Introduction

The automatic detection and classification of epileptic EEG have beneficial effects on the diagnosis and therapy of epilepsy. The EEG, which contains important information about the conditions and functions of the brain, is fundamental for diagnosing brain nervous diseases and useful for both physiological research and medical applications.

Automated epileptic seizure detection methods, based on the analysis of EEG recordings, have been developed with mixed results. For the scalp EEG, the applications for epileptic seizure detection included automatic labeling of the seizure in the EEG for faster reporting [1]. This approach had a difficult task to give many artifacts, which meant that it was not widely applied to epileptic seizure detection in clinical practice. The seizure detectors, in general, had two classes, including direct-feature-based seizure detectors [2-3] and complex seizure detectors involving classification

* Corresponding author.

algorithms [4-5]. The former class is typically more transparent and easier to understand than latter class.

There is growing evidence that the electrical activities of the brain, which showed by the EEG, are complex nonlinear dynamic systems. Considering this, the nonlinear dynamics theory may be a better approach than traditional linear methods in characterizing the nature of the electrical activities of the brain [6-8], and some methods have been proposed recently. In reference [6], under different mental states, the nonlinear parameters were evaluated from the EEG signals: correlation dimension, largest Lyapunov exponent, Hurst exponent and approximate entropy. Reference [7] analyzed largest Lyapunov exponent and the approximate entropy before and after the epileptic seizure. Reference [8] applied the local projective nonlinear noise reduction method to reduce the noise in the EEG signals.

Recently, a bridge between nonlinear time series analysis and complex networks theory has been built. Zhang and Small [9] first proposed a transformation from pseudoperiodic time series to complex networks, and the dynamics of nonlinear time series were encoded into the topology of the corresponding networks. Lacasa et al. [10] showed an alternative algorithm (the visibility algorithm) to convert a time series into a graph. Several networks topology statistical characteristics have been discussed in the reference [11], such as the joint degree distribution and the betweenness centrality, which were capable of providing a comprehensive statistical characterization of the dynamics from different time series. Reference [12] illustrated the potential of the complex networks, built by the recurrence matrix of the time series, for the detection of dynamical transitions. In Reference [13], Gao et al. associated different aspects of time series dynamics with the topological indices of the directed weighted complex networks. Considerable attention has been devoted to the nonlinear time series analysis based on complex networks theory, which provided a novel perspective for series dynamics analysis.

The classification accuracy can be improved by having a better understanding of the EEG signals. In this study, we propose and employ the Shannon entropy of the degree distribution of the resulting networks (NDDE) as a classification feature to automatically detect the epileptic ictal EEG signals from the EEG signals. At first, the EEG series is transformed to complex networks. Then the degree distribution (DDF) of the resulting networks is explored and the NDDE is used as the classification feature. The experiment results show that the shapes of DDFs are different and the classification accuracy reaches the maximum value 96.25%.

2 The Feature Based on the Statistical Property of the Networks Transforming from Series

The statistical properties of complex networks theory provide us with an effective tool for understanding the time series dynamics in a global way and excavate information that is not available from classical nonlinear time series analysis methods. For constructing the complex networks from the nonlinear time series, the series is divided up into individual cycles and each cycle is then treated as a node in the complex networks.

A time series denoted as $\{s_1, s_2, \dots, s_j, \dots, s_m\} (j \in [1, m])$, where the s_j is the j_{th} point in the time series and the length of series is m . The time series is separated into several non-overlapping and equal-length (l length each) cycles, which read

$$\{C_i = (s_{(i-1) \times l + 1}, s_{(i-1) \times l + 2}, \dots, s_{(i-1) \times l + l}) \mid i = 1, 2, \dots, N\}. \tag{1}$$

All cycles are denoted as $\{C_1, C_2, \dots, C_N\}$. Sometimes the last few points of series need to be abandoned to ensure that the product of N and l must less than $m (N \times l \leq m)$.

Regarding each cycle as a point in l -dimensional space, d_{ij} is used for describing the distance between the two cycles C_i and C_j . The distances of every pair-wise cycle are measured by either Euclidean distance or linear correlation coefficient. The Euclidean distance between C_i and C_j is selected, considering its physically meaningful and easily understanding, and is defined as

$$d_{ij} = \frac{1}{l} \times \sqrt{\sum_{k=1}^{k=l} (C_i(k) - C_j(k))^2}. \tag{2}$$

$C_i(k)$ is the k_{th} point of the cycle C_i . The phase space distance d_{ij} characterizes the high-dimensional distance between cycles, since two cycles with a smaller distance will be close in phase space. It is a pair-wise test of all phase space vectors among each other, whether or not they are close. Then a square, symmetric distance matrix is obtained, denoted as $D = (d_{ij})_{N \times N}$.

The idea is that two nodes are destined to connect if the phase space distance between the corresponding cycles is less than a predetermined value ε . The distance matrix D can be converted into a new matrix $A = (a_{ij})_{N \times N}$, called adjacency matrix, with choosing a critical value ε . The rules of converting read

$$a_{ij} = \begin{cases} 1 & (d_{ij} \leq \varepsilon) \\ 0 & (d_{ij} > \varepsilon) \end{cases}. \tag{3}$$

All the points and their connections form a network, which completely can be represented by the matrix A . The conditions $a_{ij}=1$ and $a_{ij}=0$ correspond to connection and disconnection, respectively. The topological structure of this network can be described with the adjacency matrix.

The value of ε , which regards as the filter parameter, determines the resulting network's characteristics. The nodes in pairs in the network are also connected with weak correlations, with extremely large ε , which result in the time series' physically meaningful correlations submerged by noises (considering noises are weak correlations). With the value of ε decreasing, the number of connections becomes smaller and smaller among the network points. More and more noises are filtered out. However, if the critical value becomes extremely small, some of the physically meaningful connections are also filtered out and the number of connections decreases rapidly,

which may lead to strong statistical fluctuations due to a small finite number of connections. Consequently, we can expect an interval of the critical value, $\varepsilon \in (\min(d_{ij}), \max(d_{ij}))$, in which the topological structure of networks can capture the dynamic characteristics of the time series.

Because no complete theory can guide us to find out the interval of ε , a proper value of ε can be found just by simulating a special dynamical process of changing the value of ε , i.e., changing the value of ε monotonously while keeping the other parameters unchanged and testing the classification accuracy in every situation. If a wide interval of ε exists, in which the DDF of the resulting networks have almost the same characteristic, we can conclude that the dynamical process is governed by the same law and use this characteristic to classify the time series with the different dynamics.

After the complex networks is constructed from the nonlinear time series, the basic statistical property DDF, which denoted as $P(k)$ and defined as the probability that a node chosen uniformly at random has degree k , is investigated. The degree k of a node is the number of the nodes directly connected with it. The DDF describes the heterogeneous properties and consequently can characterize the different dynamics of the networks.

The different shapes of DDF can be used as a classification feature to discriminate the data, because the shapes of the DDF of the networks corresponding to time series with different dynamics are differentiable. Since the shape of the DDF is not easily to quantify, it is inappropriate to take it as a classification feature.

Entropy is a state function of the system and directly reflects the uniformity and the condition of the system. In statistical mechanics, entropy is a measure of the number of ways in which a system may be arranged, often taken to be a measure of ‘disorder’ (the higher the entropy, the higher the disorder). Here we use the Shannon entropy to measure the ‘disorder’ of the DDF and the NDDE is selected as a feature to classify the time series with different dynamics, defined as

$$NDDE = - \sum_{i=0}^{i=\max(k)} P(i) \times \log_2 P(i). \quad (4)$$

The value of NDDE is small when the distribution has several big probability values. However, the NDDE value is large, when the distribution has a lot of small probability values (the emergence of every degree value has more average probability).

3 Experiment Results and Analysis

The performance of the proposed approach is illustrated by testing the EEG data, which obtained from the six clinical diagnosed epilepsy cases in the Qilu Hospital of Shandong University. The epileptic EEG data file contained 100 ictal EEG data (S001~S100) and 100 interictal EEG data (F001~F100). Every EEG datum was sampled at a rate of 128 Hz, and had 4096 points. For the purpose of comparison, 200 segments of ictal EEGs and 200 segments of interictal EEGs are taken out from the epileptic EEG data file (every datum in the file is divided into two equal-length segments of 2048 points).

In this study, the cycle length l is selected as 8. Here, considering the ictal EEGs as the positive class, and the interictal EEGs as the negative class.

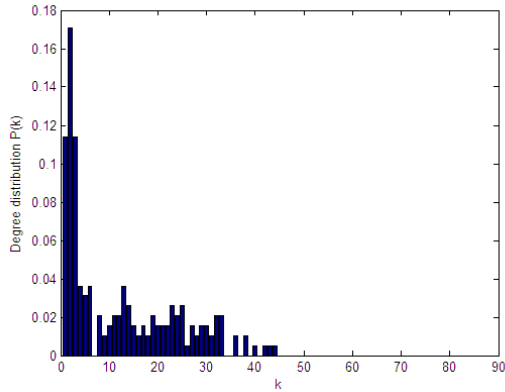


Fig. 1. Degree distribution of the networks transforming from an ictal EEG

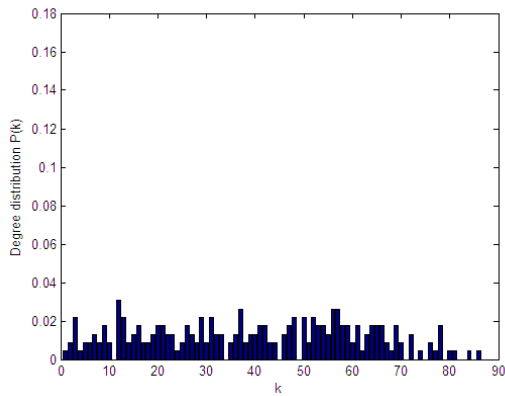


Fig. 2. Degree distribution of the networks transforming from an interictal EEG

Figure 1 plots $P(k)$ for the networks transforming from an ictal EEG signal, and Figure 2 plots $P(k)$ for the networks transforming from an interictal EEG signal. It is found that the shape of the DDF of the ictal EEG signal is different from the shape of the DDF of the interictal EEG signal. In Figure 1, the distribution only has value in the region $k \in [0,45]$. The maximum value is 0.175 and the minimum value is 0. In Figure 2, the distribution has value in the region $k \in [0,90]$, and the values of 0.04 and 0 are its maximum and minimum value, respectively. The DDF of an ictal EEG signal has a fewer numbers of values but a larger fluctuation range than an interictal EEG signal's DDF. The similar results are obtained in the other DDFs.

The value of filter parameter ε , determines the resulting network's characteristics, and therefore affects the classification accuracy. In order to find out the best performance

Table 1. The accuracy of the approach under different values of ε

ε	Ictal EEG		Interictal EEG		Accuracy (%)
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	
6	89.00	97.50	97.50	89.00	93.25
7	92.50	95.00	95.00	92.50	93.75
8	93.00	96.00	96.00	93.00	94.50
9	91.50	98.00	98.00	91.50	94.75
10	92.00	98.00	98.00	92.00	95.00
11	95.50	95.50	95.50	95.50	95.50
12	98.00	94.00	94.00	98.00	96.00
13	98.00	94.50	94.50	98.00	96.25
14	99.00	93.50	93.50	99.00	96.25
15	98.50	90.50	90.50	98.50	94.50

of the classification method based on NDDE the value of ε is focused on and the accuracies with the value of ε changing from 6 to 15 are showed in the Table 1.

With the ε increasing from 6 to 14, the correct classification rate (Accuracy) of the approach increases monotonously from 93.25% to 96.25%. The classification accuracy always higher than 93.25%, when the value of ε changing from 6 to 15. The maximum value of accuracy (96.25%) obtained when the value of ε is 13 or 14.

Figure 3 plots the best classification result showing in the Table 1, when ε is 13. There are two hundred ‘*’ points and two hundred ‘o’ points representing the NDDE of the interictal EEGs and the ictal EEGs’, respectively. As can be clearly observed from the Fig. 3, the points ‘*’ are higher than the points ‘o’ except for a few points. Choosing a constant classification threshold as 5.6693, represented by the dotted line in the Fig. 3, the classification accuracy is 96.25%.

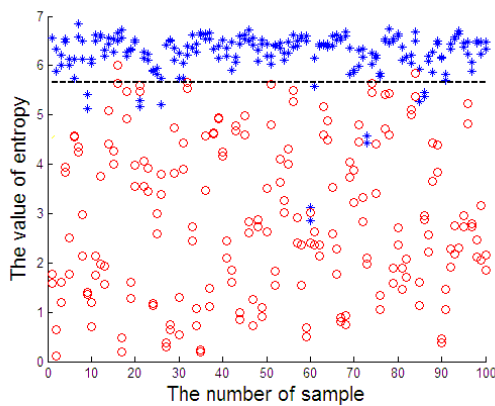


Fig. 3. The best classification result of the approach when ε is 13

4 Conclusion

The dynamics under different states of EEG signals have the essential difference. Recently, the nonlinear dynamical methods have been widely applied to analyze EEG signals. With the increasing attention to nonlinear time series dynamics analysis methods based on the complex networks theory, this study applies the DDF of the complex networks to analyze the epileptic EEG and employs the NDDE of epileptic EEG signals as a feature to automatically detect the ictal EEG signals from the epileptic EEG signals.

The results show that the shapes of DDFs are different due to the different dynamics of time series. The classification accuracies are higher than 93.25% with ε changing from 6 to 15 using the extracted feature. When the ε is 13 or 14, the classification accuracy reaches the maximum value 96.25%.

Application to epileptic EEG data demonstrates that the NDDE of EEG signal is a reliable and well-performance feature for the epileptic automatic detection.

Acknowledgement. Project supported by the National Natural Science Foundation of China (Grant No. 61201428, 61070130), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2010FQ020), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China (Grant No. BS2009SW003), and the China Postdoctoral Science Foundation (Grant No. 20100470081), the Shandong Province Higher Educational Science and Technology Program (Grant No. J11LF02).

References

1. Saab, M., Gotman, J.: A System to Detect the Onset of Epileptic Seizures in Scalp EEG. *Clin. Neurophysiol.* 116, 427–442 (2005)
2. Khan, Y., Gotman, J.: Wavelet-Based Automatic Seizure Detection in Intracerebral Electroencephalogram. *Clin. Neurophysiol.* 114(5), 898–908 (2003)
3. Haas, S., Frei, M., Osorio, I.: Strategies for Adapting Automated Seizure Detection Algorithms. *Med. Eng. Phys.* 29, 895–909 (2007)
4. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Mixedband Wavelet-Chaos-Neural Network Methodology for Epilepsy and Epileptic Seizure Detection. *IEEE Trans. Biomed. Eng.* 54(9), 1545–1551 (2007)
5. Srinivasan, V., Eswaran, C., Sriraam, N.: Approximate Entropy-Based Epileptic EEG Detection Using Artificial Neural Networks. *IEEE Trans. Biomed. Eng.* 11(3), 288–295 (2007)
6. Natarajan, K., Acharya, R., Alias, F., Tiboleng, T., Puthusserypady, S.K.: Nonlinear Analysis of EEG Signals At Different Mental States. *BioMed. Engin. OnLine* 3, 1 (2004)
7. Xie, Y., Xu, J.-X., Yang, H.-J., Hu, S.-J.: Phase-Space Reconstruction of ECoG Time Sequences and Extraction of Nonlinear Characteristic Quantities. *Acta Phys. Sin.* 51(2), 205–214 (2002)
8. Xie, Y., Xu, J.-X., Kang, Y.-M., Yang, H.-J., Hu, S.-J.: Nonlinear Noise Reduction for Electrooculograms. *Acta Phys. Sin.* 52(5), 1121–1126 (2003)

9. Zhang, J., Small, M.: Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics. *Phys. Rev. Lett.* 96, 238701 (2006)
10. Lucas, L., Bartolo, L., Fernando, B., Jordi, L., Juan, C.N.: From Time Series to Complex Networks: the Visibility Graph. *PNAS* 105(13), 4972–4975 (2008)
11. Zhang, J., Sun, J.F., Luo, X., Zhang, K., Nakamura, T., Small, M.: Characterizing Pseudoperiodic Time Series through the Complex Network Approach. *Physica D* 237, 2856–2865 (2008)
12. Norbert, M., Jonathan, F.D., Zou, Y., Reik, V.D., Jurgen, K.: Complex Network Approach for Recurrence Analysis of Time Series. *Physics Letters A* 373, 4246–4254 (2009)
13. Gao, Z.K., Jin, N.D.: A Directed Weighted Complex Network for Characterizing Chaotic Dynamics from Time Series. *Nonlinear Analysis: Real World Applications* 13, 947–952 (2012)

Threshold Estimation Method for Spectrum Sensing Using Bootstrap Technique

Liping Luo, Wei Zhou, and Huazhi Meng

School of Information Science and Engineering, Guangxi University for Nationalities,
Nanning, China
lping.luo@gmail.com

Abstract. In this paper, the bootstrap technique is applied to spectrum sensing for cognitive radio networks. A novel test threshold estimation method based on bootstrap is proposed. From the simulation results, it is seen that the proposed bootstrap procedure can provide satisfied detection performance while only requires the smallest samples compared with the existing methods. Therefore, the proposed method is very accurate and efficient for spectrum sensing.

Keywords: Bootstrap techniques, cognitive radio, spectrum sensing, test threshold.

1 Introduction

Cognitive radio [1] has been considered as a promising technique to improve the spectrum utilization for future wireless systems. In cognitive radio networks, spectrum sensing plays a fundamental role and a variety of methods have been proposed recently in [2]-[6] to detect the presence of primary users. Among them, energy detection [4] requires no information of the primary signal thus attracts much attention in spectrum sensing. However, it requires the noise variance to set the test threshold and the performance deteriorates significantly due to the noise uncertainty. To circumvent this shortcoming, some eigenvalue-based methods are developed in [5],[6]. They do not need any priori information of the primary signal and the noise variance, thus are more robust than energy detection. In eigenvalue-based algorithms, the test threshold can be obtained by theoretical formula of the false alarm probability (P_f) or Monte-Carlo simulations. However, the expression of P_f is asymptotical and only holds for large M, N . When M, N are small, the calculation through P_f is not accurate. Although Monte-Carlo simulations can provide accurate estimation of the test threshold, it needs at least 10^4 random experiments. Thus, large samples and computations are required for the existing threshold estimation methods, which limits the efficiency of the eigenvalue-based algorithms.

The bootstrap is a statistical method which is originally introduced by Bradley Efron [7]. With the bootstrap, the original data is reused, thus a new set of experiments is not needed. Specifically, the original observations are randomly reassigned and the estimate is computed. These assignments and recomputations are

done for many times and considered as repeated experiments. Hence, real repeated experiments are not required, which is the main advantage of the bootstrap over Monte-Carlo simulations. In [8]-[10], Zoubir reviewed the applications of the bootstrap technique in the signal processing field, especially for signal detection [11], the bootstrap exhibits good detection performance even for small sample sizes.

In this paper, the bootstrap technique is applied to spectrum sensing. The assumption of large sample sizes is removed. The test threshold for a given P_f is estimated using bootstrap. Simulation results show that the threshold determined by the proposed bootstrap procedure is almost the same as the value by Monte-Carlo simulations, but small samples and no repeated experiments are exploited in bootstrap procedure.

2 Signal Model and Eigenvalue-Based Detection

2.1 Signal Model

As in [6], the secondary user is deployed with M antennas to detect the primary signal. The spectrum sensing problem can be formulated as the following binary hypothesis test:

$$\begin{aligned} \mathcal{H}_0 : \quad & \mathbf{x}(n) = \mathbf{v}(n), \quad n = 0, \dots, N - 1 \\ \mathcal{H}_1 : \quad & \mathbf{x}(n) = \mathbf{h}s(n) + \mathbf{v}(n), \quad n = 0, \dots, N - 1 \end{aligned} \quad (1)$$

where $s(n)$ is the n th sample of the primary signal, which is assumed to be quadratic phase shift keying (QPSK) signal but the modulation mode is unknown to the secondary user. $\mathbf{v}(n) \in C^{M \times 1}$ is the complex Gaussian noise with unknown variance. $\mathbf{h} \in C^{M \times 1}$ is the channel coefficient between the primary transmitter and the secondary receiver which is assumed to be constant during the sensing interval. $\mathbf{x}(n) \in C^{M \times 1}$ is n th sample of the received signal, all of the received samples can be formulated as an $M \times N$ -dimension signal matrix $\mathbf{X} = [\mathbf{x}(0), \mathbf{x}(1), \dots, \mathbf{x}(N - 1)]$.

2.2 Eigenvalue-Based Detection

Without priori knowledge of the noise variance and the primary signal, some eigenvalue-based detectors, such as maximum-to-minimum ratio eigenvalue (MME), arithmetic-to-geometric mean (AGM) and generalized likelihood ratio test (GLRT), are proposed in [5][6]. Using the eigenvalues of the sample covariance matrix, the detectors are developed as follows:

$$T_{MME} = \frac{\hat{\lambda}_1}{\hat{\lambda}_M} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma \quad T_{AGM} = \frac{\frac{1}{M} \sum_{m=1}^M \hat{\lambda}_m}{\left(\prod_{m=1}^M \hat{\lambda}_m \right)^{1/M}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma \quad T_{GLRT} = \frac{\hat{\lambda}_1}{\sum_{m=1}^M \hat{\lambda}_m} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\geq}} \gamma \quad (2)$$

where $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_m \geq \dots \geq \hat{\lambda}_M$ is the ordered eigenvalues of the sample covariance matrix $\hat{\mathbf{R}}_x = \frac{1}{N} \mathbf{X} \mathbf{X}^H$. γ is the test threshold, which is dependent on the

behavior of eigenvalues under \mathcal{H}_0 . For a given false alarm probability P_f , there are two methods to determine the value of γ . One is Monte-Carlo simulations, which need at least 10^4 random experiments. The other is theoretical calculation, taking the GLRT detector for an example, γ is asymptotically calculated by

$$\gamma = \frac{\left(1 + \sqrt{\frac{M}{N}}\right)^2}{M-1} + \frac{1}{N(M-1)} \frac{(\sqrt{N} + \sqrt{M})^{4/3}}{(\sqrt{MN})^{1/3}} F_{TW2}^{-1}(1 - P_f) \quad (3)$$

where $F_{TW2}(\cdot)$ is the second-order Tracy-Widom distribution function [12], the value can be obtained through a lookup table. However, the theoretical value only holds for large sample size.

3 Spectrum Sensing Using Bootstrap

As above mentioned, the asymptotical expression of the false alarm probability is not valid for small sample size. Thus, the test threshold calculated according to (3) may not be accurate, which will degrade the detection performance. To solve this problem, the bootstrap is applied to spectrum sensing, a procedure using bootstrap is proposed to estimate the test threshold for eigenvalue-based detection when the sample size is small.

Denote the received samples \mathbf{X} as original data set. Different from the Monte-Carlo simulations, the bootstrap procedure runs based on the original data set, rather than the repetition of the real experiment. The first step in bootstrap procedure is to randomly select a column from \mathbf{X} to generate new random samples, then repeat the first step for N times, a bootstrap data set \mathbf{X}^* with the same size as \mathbf{X} is generated.

$$\mathbf{X}^* = [\mathbf{x}(0)^*, \dots, \mathbf{x}(N-1)^*] \quad (4)$$

The sample eigenvalues of the bootstrap data set \mathbf{X}^* are

$$\lambda_1^* \geq \dots \geq \lambda_m^* \geq \dots \geq \lambda_M^* \quad (5)$$

From the bootstrap eigenvalues, the required test statistic defined in (2) can be formed. Repeating the procedure B times gives an estimate of the distribution of this test statistic which can be used for detection. The bootstrap procedure for threshold estimation is shown in Table 1.

Applied the bootstrap threshold estimation procedure to the eigenvalue-based detection, an accurate threshold value can be obtained. Comparing the test statistic with the threshold, the secondary user can decide whether the primary signal is present or not. Since the proposed method does not require large samples, the efficiency of the sensing method can be improved.

4 Simulation Results

The simulation parameters are set as follows [6]: the secondary receiver is deployed with $M = 4$ antennas, the number of the observed samples is $N = 10$, independent

Table 1. Test threshold estimation algorithm using bootstrap

Step 1.	Experiment. Receive the observed samples, form $M \times N$ matrix, $\mathbf{X} = [\mathbf{x}(0), \dots, \mathbf{x}(N - 1)]$;
Step 2.	Resampling. Randomly select a column from \mathbf{X} with replacement, repeat N times to form the bootstrap data: $\mathbf{X}^* = [\mathbf{x}^*(0), \dots, \mathbf{x}^*(N - 1)]$;
Step 3.	Centering. Center \mathbf{X}^* by subtracting the sample mean from each column, $\mathbf{x}^*(i) \leftarrow \mathbf{x}^*(i) - \frac{1}{N} \sum_{j=0}^{N-1} x^*(j), i = 0, \dots, N - 1$;
Step 4.	Compute the sample covariance matrix, $\hat{\mathbf{R}}^* = \frac{1}{N} \mathbf{X}^* \mathbf{X}^{*H}$;
Step 5.	Perform singular value decomposition (SVD) on $\hat{\mathbf{R}}^*$, obtain the sample eigenvalues: $\lambda_1^*, \dots, \lambda_M^*$;
Step 6.	Compute the test statistic: $T_{GLRT}^* = \frac{\lambda_1^*}{\sum_{m=1}^M \lambda_m^*}$;
Step 7.	Repeat the Step 2~6 for B times to obtain B Bootstrap statistics: $T_{GLRT}^*(1), \dots, T_{GLRT}^*(b), \dots, T_{GLRT}^*(B)$;
Step 8.	Ranking. Rank $T_{GLRT}^*(b)$ into decreasing order, the q th value $q = \lceil P_f \times B \rceil$ of T_{GLRT}^* is the test threshold, i.e., $\gamma_{boot} = T_{GLRT}^*(q)$.

Rayleigh fading channels are simulated between the primary transmitter and the secondary receiver. Three threshold calculation methods, including theoretical formula of P_f , Monte-Carlo simulations and the proposed bootstrap procedure are employed for GLRT spectrum sensing method. The simulation times and the bootstrap resampling times are $B = 1000$.

For a given false alarm probability, the values of test threshold determined by the abovementioned methods are listed in Table 2. Since the real experiment are repeated for many times in Monte-Carlo simulations, the obtained threshold is relatively accurate, then is regarded as a reference value for comparison.

Table 2. Test threshold calculated by three methods

P_f	γ_{theo}	γ_{boot}	γ_{sim}
0.01	0.93	1.7261	1.6023
0.05	0.8683	1.2937	1.3047
0.1	0.8361	1.1713	1.1734

From Table 2, it is seen that the threshold obtained by theoretical formula deviates significantly from the reference value. The reason is that the number of effective samples is $MN = 40$, which is very small. Thus, the expression of P_f does not hold, naturally, the derived threshold is not accurate. On the other hand, the threshold obtained by bootstrap is almost the same as the reference value. However, the effective samples used in bootstrap is $MN = 40$, it is much less than that used in Monte-Carlo simulations which need $MNB = 40000$ effective samples. Moreover, the bootstrap does not require to repeat the experiment.

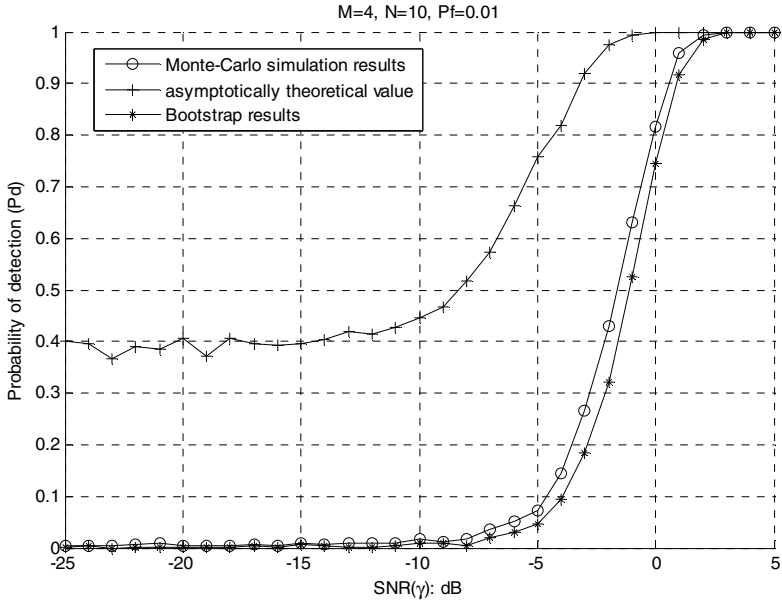


Fig. 1. Pd v.s. SNR under Pf = 0.01

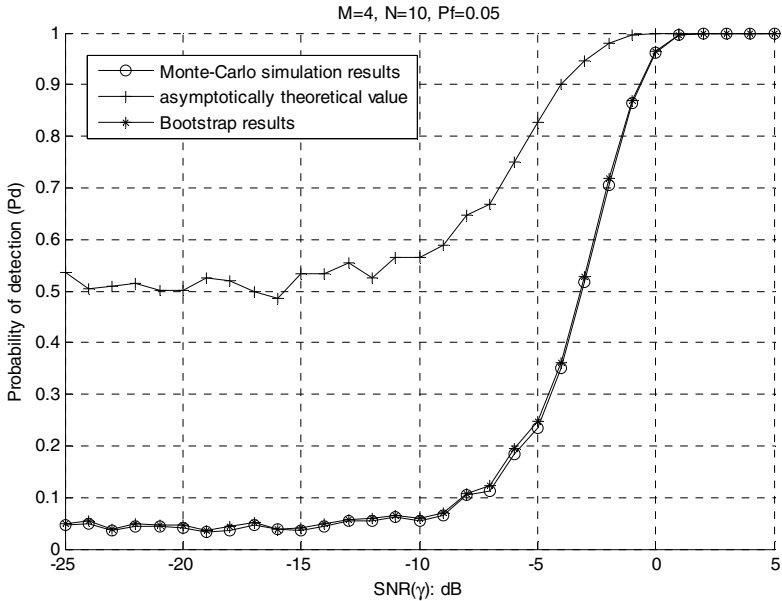


Fig. 2. Pd v.s. SNR under Pf = 0.05

The detection performance of GLRT of which the threshold is estimated by the three methods are shown in Fig.1-2. It is observed that the detection probability of theoretical calculation deviates much from the simulation value. The reason is that the

threshold obtained by theoretical formula is less than the real threshold, thus it causes larger detection probability, but it is not correct. On the other hand, the detection performance of bootstrap with 40 effective samples matches very well with that obtained by Monte-Carlo simulations with 40000 samples, especially at $P_f = 0.05$, which indicates the accuracy and efficiency of the bootstrap method.

5 Conclusions

Spectrum sensing is a fundamental task in cognitive radio. Recently, some eigenvalue-based detection methods, such as GLRT, have been proposed for secondary user. For these eigenvalue-based approaches, theoretical calculation or Monte-Carlo simulations are employed to determine the test threshold. However, large samples are required for accurate estimation, which degrades the efficiency of the approaches. To overcome this problem, the bootstrap technique is applied to develop a test threshold estimation procedure for spectrum sensing. Compared with the existing methods, the proposed bootstrap procedure can provide good detection performance with small sample size.

Acknowledgments. This research was supported by Guangxi Natural Science Foundation (No.2012GXNSFBA053162), the Bring in Talents Projects of Guangxi University for Nationalities (No.2012QD018, No.2009QD022).

References

1. Haykin, S.: Cognitive Radio: Brain-empowered Wireless Communications. *IEEE J. Sel. Areas Commun.* 23(2), 201–220 (2005)
2. Cabric, D., Tkachenko, A., Brodersen, R.W.: Spectrum Sensing Measurements of Pilot, Energy, and Collaborative Detection. In: *Proc. MILCOM*, pp. 1–7. IEEE Press, Washington, DC (2006)
3. Sutton, P.D., Nolan, K.E., Doyle, L.E.: Cyclostationary Signatures in Practical Cognitive Radio Applications. *IEEE J. Sel. Areas Commun.* 26(1), 13–24 (2008)
4. Digham, F.F., Alouini, M.S., Simon, M.K.: On the Energy Detection of Unknown Signals over Fading Channels. *IEEE Trans. Commun.* 55(1), 21–24 (2007)
5. Zeng, Y., Liang, Y.C.: Maximum-Minimum Eigenvalue Detection for Cognitive Radio. In: *Proc. IEEE 18th Int. Symp. Pers., Indoor, Mobile Radio Commun.*, pp. 1–5. IEEE Press, Athens (2007)
6. Wang, P., Fang, J., Han, N., Li, H.: Multiantenna-Assisted Spectrum Sensing for Cognitive Radio. *IEEE Trans. Vehicular Tech.* 59(4), 1791–1800 (2010)
7. Efron, B.: *Computers and the Theory of Statistics: Thinking the Unthinkable*. SIAM Review (1979)
8. Zoubir, A.M.: The Bootstrap: Signal Processing Applications. *IEEE Signal Process. Mag.* 15(1), 56–76 (1998)
9. Zoubir, A.M.: The Bootstrap: a Powerful Tool for Statistical Signal Processing with Small Sample Sets. In: *Proc. of ICASSP* (1999)
10. Zoubir, A.M., Iskander, D.R.: *Bootstrap Techniques for Signal Processing*. Cambridge University Press, New York (2004)
11. Brcich, R.F., Zoubir, A.M.: Detection of Sources Using Bootstrap Techniques. *IEEE Trans. Signal Process.* 50(2), 206–215 (2002)
12. Bejan, A.: Largest eigenvalues and sample covariance matrices, Tracy-Widom and Painleve II: computational aspects and realization in SPlus with applications. Preprint (2005), <http://www.vitrum.md/andrew/MScWrwck/TwinSplus.pdf>

Palm Print Feature Extraction and Recognition Based on BEMD-ICA II and LS-SVM

Gui-Ping Dai

Department of Electronic & Information Engineering, Suzhou Vocational University,
Suzhou 215104, Jiangsu, China
luaiping_0127@163.com

Abstract. A novel palm print recognition method based on BEMD (Bidimensional Empirical Mode Decomposition) and ICA II (Independent Component Analysis) integrated with LS-SVM was proposed. Firstly, BEMD was adopted to decompose the preprocessed palm print image into a group of IMFs (Intrinsic Mode Functions), and then a fast fixed-point algorithm of ICA II (Fast ICA II) was carried out to extract the palm print feature subspaces of IMF sub images matrix, before which PCA (Principal Component Analysis) was used to decrease the dimensions of input images matrix; Finally, by means of LS-SVM classification algorithm, the recognition performance of the integrated method (BEMD+ICA II) was tested on the Hong Kong Polytechnic University palm print database. Simulation results show that compared with ICA II, BEMD-ICA II method can more effectively and accurately extract the palm print features, so that can achieve the superior SNR (signal-noise-ratio) of reconstructed image. Moreover, compared with the Standard Euclidean distance classification method, LS-SVM can achieve the higher recognition rate.

Keywords: BEMD (Bi-dimensional empirical mode decomposition), ICA (Independent component analysis), LS-SVM (Least square support vector machine), Palm print feature extraction, Palm print mode recognition.

1 Introduction

Palm print recognition is a new recognition technology in the field of biological authentication, which is applied to the identity automatic confirmation according to the palm texture characteristics including palm surface flexor line, wrinkles, nipple grain and so on [1]. The process is roughly divided into four parts: image collection, pre-treatment, feature extraction, classification decision-making.

And the accurate feature extraction of the palm print is the key of the rapid and accurate recognition. Moreover, as for the palm print images, its most important feature information (such as edges, details, etc.) has close relation with the high-order statistical properties between pixels, namely the higher-order statistical properties usually contain the important structural and phase information of image. Therefore, ICA algorithm based on the high-order statistical properties has unique advantages in the palm

print feature extraction. It is based on the statistical independence principle. By means of the optimization algorithms, it can make the transformed components not only irrelevant, but also statistical independent, thus can more fully reveals the essential characteristics of image[2,3]. On the other hand, BEMD is an adaptive time-frequency localized and multi-scale analysis method, which can adaptively decompose an image into a group of IMF (Intrinsic Mode Function) sub images with different intrinsic time scales according to the image inherent characteristics [4]. Each of IMFs not only can reveal the intrinsic real physical information of image, but also can reduce the interference or coupling between the image characteristic information, and thus provides a more effective and novel method for improvement of the palm print recognition rate.

On the other hand, SVM is a new machine learning algorithm proposed by Vapnik etc. according to statistical theory. It is based on the VC (Vapnik Chervonenks) dimension theory and the SRMI (structural risk minimization inductive) principle. When there is only a small sample set available, it has a simple structure, fast learning speed and can converge to the global optimum. Therefore, it has become a hotspot in the field of time series prediction and pattern recognition[5,6].

In this paper, combining with the adaptive decomposition of BEMD, higher-order statistical independency of ICA and higher recognition rate of LS-SVM, a novel palm print feature extraction and recognition method based on BEMD-ICA II and LS-SVM was proposed.

2 Basic Principle

2.1 BEMD Principle

BEMD algorithm is the expansion of one-dimensional EMD, which can extract the details of image on different characteristic scales through the auto-merging sifting process. For the two-dimensional image $f(x, y)$, $x = 1, \dots, M$, $y = 1, \dots, N$, the BEMD decomposition with N layers of image $f(x, y)$ can be described as follows[7]:

$$f(x, y) = \sum_{i=1}^N I_{imfi}(x, y) + r_N(x, y). \quad (1)$$

In which $I_{imfi}(x, y)$ stands for the decomposed i th IMF sub image and $r_N(x, y)$ stands for the final residual component. The key point of BEMD algorithm lies in the search algorithm for local extreme points of image and the envelopment surface fitting algorithm, especially the latter one, namely that is extreme value point interpolation algorithm. Thereby, BEMD adaptively breaks down the image $f(x, y)$ into N stationary IMF sub images and the residual component according to the image inherent characteristics and each IMF corresponds to a unique intrinsic time scale. Therefore, the palm print features can be represented under different resolutions and the interference or coupling between the image characteristic information can be reduced, so that the palm feature can be extracted effectively.

2.2 Statistics Independent Feature Extraction of Image Based on ICA II

ICA adopted to extract the palm print features has two kinds of structures: ICA I and ICA II . In ICA I ,the extracted base images are statistical independent, while feature coefficients in the base image subspace are not independent. However, the purpose of ICA II is to extract the statistics independent feature coefficients of input image. In ICA II , column vectors of the training image matrix \widehat{X}_{train} stand for the different sub image blocks which are observed values, and row vectors indicate pixels which are random variable. Assume W_I is the transition matrix which can be calculated by Fast-ICA learning algorithm based on the biggest negative entropy criterion. Define $A = W_I^{-1}$ as the feature base image matrix, each column vector of which corresponds to a base image. Then denote $B = W_I(\widehat{X}_{train})^T$ and each column vector of B corresponds to the statistics independent feature coefficients of a base image.

Assume the dimension of training image matrix \widehat{X}_{train} is $p \times n$, in which p is number of image pixels and n is number of sub image blocks. In order to reduce the computational complexity of ICA II , PCA algorithm is adopted to remove the second-order statistical characteristics relevance to decrease the dimension of the training image matrix \widehat{X}_{train} . Suppose V_k is a matrix with dimension of $p \times k$ which consists of the former k eigenvectors, where p is the number of image pixels and k is the number of principle components. As a result, the PCA coefficients of \widehat{X}_{train} which has been carried out the zero mean and whitening processing can be obtained:

$$R_k = \widehat{X}_{train} V_k . \tag{2}$$

When the estimated value of \widehat{X}_{train} with the least mean square error can be defined as $\widehat{X}_{train} = R_k V_k^T$. And then denote V_k^T as the input of ICA II , so that the statistical independent ICA II coefficients of \widehat{X}_{train} can be described as follows:

$$B = W_I R_k^T . \tag{3}$$

As a result, the statistics independent feature coefficients of base images can be extracted effectively which corresponds to the column vectors of B . While each column vector of $A = W_I^{-1}$ corresponds to a base image.

2.3 LS-SVM Recognition Algorithm

The basic idea behind the Least Square Support Vector Machine (LS-SVM) classifying algorithm is as follows: the input space is transformed into a high dimensional space through the nonlinear transform defined by kernel functions. The sample data can be thus be separated linearly. The optimal classifying plane can be constructed in the high dimensional space so that the decision rules for the sample classification can be formulated.

The LS-SVM theory caters mostly to the 2-classification case. As for the multiple classification case, the basic idea is to transform the problem into a 2-classification problem, which is usually implemented using a combined classifier. Both the one-to-one and the one-to-many methods can be used in this process, of which the first is used in this paper. The algorithm constructs all the possible 2-class classifiers in the $N(N-1)/2$ LS-SVM sub-classifiers altogether. A poll is carried out to all these 2-class classifiers. And the class with the most polls is determined as the class that the point belongs to. Assume the N class training set is:

$$(x_i, y_i), i = 1, 2, \dots, n, x \in R^d, y_i \in \{1, 2, \dots, N\} . \tag{4}$$

In which n is the number of training sets, d is the dimension of each training set, x_i denotes input the training sample and y_i denotes the class number.

When the sample x is fed into the 2-class classifier constructed from the m -th class sample and the n -th class sample, the classifying function is:

$$f_{mn}(x) = \text{sign}[\sum_s (w^{mn})^T K(x_i, x) + b^{mn}] . \tag{5}$$

In which sign is the signed function, w^{mn} is the vertical vector on the classifying hyper plane, $K(x_i, x)$ is the kernel function, b^{mn} is the location of the hyper plane and s is the support vector.

If x is determined to be a member of the m -th class by $f_{mn}(x)$, one poll is added to the m -th class. Otherwise, one poll is added to the n -th class. After all the $N(N-1)/2$ 2-class classifiers have made the decision for x , the class that has the most polls is determined to be the one x belongs to.

3 Palm Print Feature Extraction Based on BEMD- ICA II

3.1 BEMD Decomposition of Palm Print Image

First of all, BEMD algorithm is carried out to break down the preprocessed palm print image on five lays, in order to obtain five IMF sub images with different characteristics scales which size is from small to large (accordingly the instantaneous frequency of IMFs is from high to low). Within this BEMD decomposition, the morphological algorithm is adopted to acquire the local extreme points of image and the CSRBF (Compactly Supported Radial Basis Function) interpolation based on the conjugate gradient method is used to carry out the envelope curve surface fitting.

According to the essence of BEMD, IMF1 with the smallest scale contains the most detailed structural and marginal information, while the remaining IMFs with larger scales reflects the basic structure and change trend information. Because the base function is obtained adaptively from the original image, therefore IMF sub images can reveal the intrinsic real physical information, namely can represent the frequency range and fluctuation of the original image in a certain feature scale.

3.2 Palm Print Feature Extraction Based on BEMD- ICA II

Palm print features extracted by means of ICA I are statistically independent feature base images which can reflect local properties in different parts of palm print. While the features extracted by ICA II are statistically independent feature coefficients of base images which can reflect the global properties of palm print and global properties include more abundant palm print information. So that this paper adopts BEMD combining with ICA II to extract the palm print features.

IMF1 ~ IMF5 sub images derived from BEMD represent distribution information in each feature scale of the original palm print image, so that IMF1 ~ IMF5 are treated as the input matrix of ICA II. A thousand sub image blocks with the size of 8×8 in pixel are randomly extracted from each IMF. Thereby, the input of ICA II is an image matrix \hat{X} with the size of 64×5000 . Furthermore, column vectors of \hat{X} indicate the different sub image blocks which are observed values, and row vectors indicate pixels which are random variable. In order to reduce the computational complexity of ICA II, PCA algorithm is adopted to remove the second-order statistical feature relevance to decrease the dimension of the input matrix \hat{X} . The key point of PCA is to select the appropriate number k of primary components. The testing experiment shows that if $k = 40$ the signal-to-noise ratio (SNR) of the reconstructed image is the biggest. While if $k > 40$ the SNR starts to decrease. So that if $k = 40$, ICA II can achieve the highest recognition rate. As a result, the dimension of the input matrix \hat{X} can be decreased from 64×5000 to 40×5000 , which can greatly reduce the computation cost.

The transition matrix W_l can be calculated by Fast-ICA learning algorithm based on the biggest negative entropy criterion. Define $A = W_l^{-1}$ as the feature base image matrix, each column vector of which corresponds to a base image. Then denote $B = W_l R_k^T$ and each column vector of B corresponds to the statistics independent feature coefficients of a base image. The feature base image matrix A with the size of 64×40 is shown in Fig.1. Then each column vector of A is transformed into a sub image block with size of 8×8 . As a result, a set including 40 feature base images is obtained which is shown in Fig.1. If 5000 sub image blocks with size of 8×8 are directly extracted from the original palm print image without BEMD decomposition, and the dimension of the input matrix is decreased into 40×5000 by means of PCA. Then ICA II is used to extract the base image set which is shown in Fig.2.

Comparing Fig.1 with Fig.2, we can see that the both ICA II base images can reflect higher-order information of the palm print, which have directionality in space and locality in the frequency domain. But compared to single ICA II, the feature base images derived from BEMD- ICA II have more obvious brightness changes and more clear marginal information, so that can more fully reflect global properties of the palm print. Furthermore, as shown in Table 1, the SNR of the reconstructed image based on BEMD- ICA II can reach 29.7709, while the SNR based on single ICA II only 24.2104. Therefore, ICA II combined with BEMD algorithm can greatly improve the precision of the palm print feature extraction and reconstructed

images. However the computation cost of BEMD-ICA II is larger than the single ICA II, the subsequent research from optimization algorithm can improve the speed, and thus can realize the real-time processing.

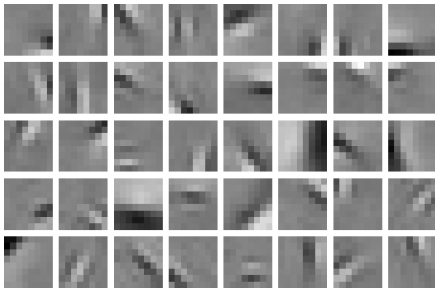


Fig.1. Base images based on BEMD- ICA II

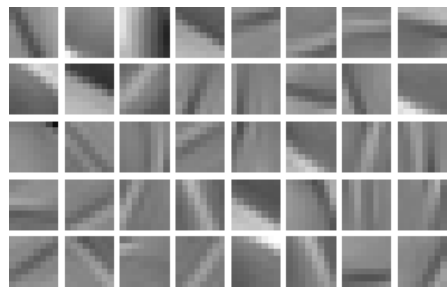


Fig.2. Base images based on ICA II

Table 1. SNR and computational cost comparison with BEMD-ICA II and single ICA II

Feature extraction methods	SNR of the reconstructed image	time-consuming cost/s
Single ICA II	24.2104	17.34
BEMD-ICA II	29.7709	142.18

4 Palm Print Recognition Simulation Based on BEMD-ICA II and LS-SVM

The palm print image analyzed in the simulation comes from the PolyU palm print image database provided by the Hong Kong polytechnic university, which includes altogether 600 images derived from 100 persons. The size of each image in pixel is 384×284. The image preprocessing method proposed in literature [8] is adopted to detect the corner, establish the coordinate system, then locate and segment the palm print image. As a result, the ROI (Region of Interest) area with size of 128×128 in pixel can be obtained.

In order to measure the recognition performance of BEMD-ICA II combined with LS-SVM, tests were respectively carried out by means of selecting different training samples and choosing different random sampling sub image blocks. Then the test results were compared with the ones by means of BEMD-ICA II combined with the standard Euclidean distance, in order to test the superiority of LS-SVM recognition performance.

Firstly, 2, 3 and 4 palm print images from six samples are randomly selected to separately train, and the corresponding remaining palm print images are used for recognition. Meanwhile, for each different training strategy, the number of sub image blocks which are randomly sampled is changed to test the influence of training

samples and sub image blocks number on BEMD-ICA II and LS-SVM. Palm print feature base images were extracted by means of the method proposed in section 3.2, and the number of primary components is 40. Moreover, LS-SVM algorithm is used for feature matching and recognition. The experiment results are shown in Table 2. We can see that the more training samples and sub image blocks, the better is the recognition performance of BEMD-ICA II .

Table 2. Recognition rates comparison with BEMD-ICA II +Standard Euclidean distance and BEMD-ICA II +LS-SVM

Algorithms	Training sam- ples number	Recognition rate/%			
		1000	3000	4000	8000
BEMD-ICA II + Standard Euclidean distance	2	95.97	96.53	96.78	97.08
BEMD-ICA II +LS-SVM	2	96.34	96.87	96.92	97.61
BEMD-ICA II + Standard Euclidean distance	3	96.78	97.36	97.54	97.93
BEMD-ICA II +LS-SVM	3	97.02	97.81	98.02	98.34
BEMD-ICA II + Standard Euclidean distance	4	97.28	97.86	98.13	98.53
BEMD-ICA II +LS-SVM	4	97.52	98.03	98.35	98.79

In order to test the superiority of LS-SVM, it has been compared with the standard Euclidean distance in the aspect of palm print recognition rates. The comparison testing results are shown in Table 2. It can be obviously seen that, compared with the standard Euclidean distance , the recognition rate of BEMD-ICA II and LS-SVM is higher, which indicates that the recognition performance based on LS-SVM is superior to the traditional distance classification. Therefore, from the two aspects of feature extraction and recognition algorithm, the palm print recognition performance has been improved.

5 Conclusion

In this paper, a novel palm print feature extraction and recognition method based on BEMD- ICA II integrated with LS-SVM was proposed. Firstly, BEMD was carried out to break down the preprocessed palm print image. Secondly, in order to reduce the computation cost, PCA algorithm was adopted to decrease the dimension of the input images matrix. And then the palm print feature subspaces of IMF sub images were extracted by a fast fixed-point algorithm of ICA II (Fast ICA II). Finally, LS-SVM was used for the palm print matching and recognition. Simulation results show that compared to single ICA II , the feature base images derived from BEMD- ICA II have more obvious brightness changes and more clear marginal information, so that

the SNR of reconstructed image is higher. On the other hand, based on the PolyU palm print image database, recognition rates of BEMD-ICA II combined with Standard Euclidean distance and BEMD-ICA II combined with LS-SVM were compared with each other by means of selecting different training samples and different random sampling sub image blocks. Recognition simulation results show that the recognition rate based on LS-SVM is higher, and thus a more effective and novel method for improving the palm print recognition performance has been provided.

References

1. Lu, X.G., Han, P.: An approach for SAR image de-noising based on 2-D wavelet transform and ICA. *Journal of Electronics & Information Technology* 30(5), 1052–1055 (2008)
2. Guo, J.Y., Yuan, W.Q.: Palm print recognition based on independent component analysis. *Opto-Electronic Engineering* 35(3), 136–139 (2008)
3. Lian, Q.S., Chen, S.Z.: Palm print identification algorithm based on energy and direction feature fusion of Gabor wavelet. *Chinese Journal of Scientific Instrument* 29(3), 556–561 (2008)
4. Song, L.X., Gao, F.J.: Compared and improved research of bidimensional empirical mode decomposition method. *Journal of Electronics & Information Technology* 30(12), 2890–2893 (2008)
5. Huan, R.H., Yang, R.L.: SAR image feature extraction and target recognition based on ICA and SVM. *Computer Engineering* 34(13), 24–28 (2008)
6. Zhou, X.H., Peng, Y.H.: Palm print recognition based on wavelet transform and support vector machine. *Computer Engineering and Application* 43(22), 238–240 (2008)
7. Wan, J., Reng, L.T.: Simulation research of applying two-dimensional empirical mode decomposition on image feature extraction. *Journal of System Simulation* 21(3), 799–801 (2009)
8. Chen, Z., Huang, L.L.: Palm print recognition based on ICA and BP neural network. *Journal of Beijing University of Aeronautics and Astronautics* 34(3), 290–293 (2008)

Detection and Classification of Defect Patterns in Optical Inspection Using Support Vector Machines

Liangjun Xie¹, Rui Huang², and Zhiqiang Cao³

¹ Houston, TX, USA

liangjunxie@gmail.com

² NEC Lab, Beijing, China

ruihuang@hust.edu.cn

³ Institute of Automation, Chinese Academy of Sciences, Beijing, China

zhiqiang.cao@ia.ac.cn

Abstract. Optical inspection techniques have been widely used in industry as they are non-destructive, efficient to achieve, easy to process, and can provide rich information on product quality. Defect patterns such as rings, semi-circles, scratches, clusters are the most common defects in the semiconductor industry. Most methods cannot identify two scale-variant or shift-variant or rotation-variant defect patterns, which in fact belong to the same failure causes. To address these problems, a new approach has been proposed in this paper to detect these defect patterns in noisy images obtained from printed circuit boards, wafers, and etc. A median filter, background removal, morphological operation, segmentation and labeling are employed in the detection stage of our method. Support vector machine (SVM) is used to identify the defect patterns which are resized. Classification results of both simulated data and real noisy raw data show the effectiveness of our method.

Keywords: optical inspection, semiconductor, pattern recognition, defect detection, support vector machine.

1 Introduction

In many industries, manual inspection cannot satisfy the modern manufacturing requirement because of the slower speed, lower inspection accuracy, and higher cost. Even the experts can only achieve about 80% accuracy in semiconductor industries [1]. Automated inspection equipment's are highly required to improve the manufacturing productivity, and the inspection accuracy. Optical methods as non-destructive methods are widely used in industry inspection. They provide not only non-contacting ways in getting information without damaging the object of interest, but also efficient means of collecting information. They are dominant in industrial inspection for products such as printed circuit boards (PCBs), semiconductor wafers, and etc. After fetching the raw data, characteristic features of the product quality are extracted, and defects are detected and classified.

Cunningham and MacKinnon [2] developed a statistical method for visual defect metrology, which includes quadrant analysis, spatial point pattern statistics and spatial

pattern recognition. Chien *et al.* [3] proposed a framework using K means clustering. However, these methods focused on the statistical analysis of defect patterns, and failed to classify meaningful defect patterns. Hsieh [4] developed a fuzzy rule-based inference system to help identify defect spatial patterns. However, it cannot identify two shift-variant or two rotation-variant defect patterns that belong to the same type. Jeong *et al.*, [5] attempted to develop a methodology to detect spatial autocorrelation and to classify defect patterns automatically based on a spatial correlogram of a wafer map. However, it does not show its robustness over the rotation-variant defect patterns.

Recently, neural networks are used as an effective means to identify defects. Non-linear high dimension features can be easily selected by neural networks and classified after the neural networks are well trained. There are two types of neural networks: supervised [6,7,15,23,24] and unsupervised. Su *et al.* [8] applied various types of supervised neural networks for the post-sawing inspection of wafers. Huang [9] used a self-supervised multilayer perceptron to detect clustered chip defects of wafers. When the defect patterns are known, supervised learning can achieve satisfactory results. Furthermore, they can transform a large number of defects into a small group of specific failure patterns and thus shorten the time and scope for troubleshooting to yield improvement.

While unsupervised learning algorithms have advantages when new defect patterns need to be added. Chen *et al.* [10] applied unsupervised learning neural networks to recognize spatial defect patterns. Lee *et al.* [11] introduced the Kohonen's Self Organizing Feature Maps (SOM) [22] as a data sampling method. Liu *et al.* [12] employed adaptive resonance theory network 1 (ART1) to recognize defect spatial patterns using wafer bin maps and claimed that ART1 was better than SOM. In their paper, only 3 defect patterns were tested, and two of them were rings. DeNicolao *et al.* [13] and Palma *et al.* [14] conducted a research with simulated data. However, both SOM and ART cannot identify two shift-variant or two rotation-variant defect patterns that in fact belong to the same failure cause [16].

To successfully recognize the defect patterns, a large number of training samples are required for neural networks, which normally takes a long time and a great effort to categorize for the training data. Multi-class support vector machines often have the advantages over neural networks [17] in such problems. Generally, neural networks suffer from multiple local minima. When they are dealing with bigger problems, overfitting is frequently observed. However, support vector machines have a global and unique solution. The computational complexity of SVMs does not depend on the dimensionality of the input space. Thus they are less prone to overfitting, and have more robust performance with respect to composite or diverse datasets. With the rapid development in support vector machine, it is widely used in many fields. Xie *et al.* [18] applied the support vector regression to enhance the image resolution. Song *et al.* [19] employed SVM for the reliability analysis of high dimensional problems. Xie *et al.* [20] systematically compared different multi-class SVMs and LVQ for control charts pattern recognition

In this paper, the most frequently observed patterns both in PCB and Wafer: ring, semi-circle, cluster, and scratches are addressed. A novel approach has been proposed to detect the defect patterns in noisy images acquired from products. In the proposed

approach, background removal, morphology operations, segmentation, and labeling have been adopted for processing the noisy data taken from micro-scope. In the classification stage, high generalization performance need to be enforced, thus supervised training networks capable of achieving good performance are more appropriate for this task [21]. Since multi-class support vector machines have advantages over neural networks with respect to the over-fitting problem, it is used as the classifier.

The remainders of this paper are organized as follows. Section 2 lists the details of the proposed approach. The methods to simulate different defect patters are described in section 3. Experiments are reported in section 4, both for simulated data and real image. Conclusions and discussions are presented in section 5.

2 Proposed Automatic Identification Algorithm

Raw images are taken from the microscope, and the defects are ranging from size of 10 pixels to 100 roughly. Usually, the defects are brighter than the background. However, the image qualities are affected by the environment, and the resolution of the camera. Lighting nearby will also have strong impact on the illumination. As a consequence, the background in one area may sometimes be brighter than the defects in other region, which makes it difficult to detect defects in such images by thresholding.

A novel defect detection and classification algorithm is proposed in this paper. Fig. 1 shows the flowchart of the method, including modules such as SVM training, testing, and the defect detection and classification. SVM training includes generating samples of the four defect patterns, extracting the features, and training the SVM using the simulated samples. Since all the defects have the same size in the training, it solves the scale-variant problem. A series of defects are generated as the training data by duplicating the same defect with some rotation, which fixes the rotation-variant issue. Shifting any irregular defect such as the semi-circle from its weighted center to the geometric center solves the shift-variant problem. It can be extended to other special defect by generating a series of training data with the same methods and adding them into the training data. SVM testing includes generating samples of the noise defect patterns, extracting the features, and testing by the trained SVM. It is used to show the performance of the trained SVM.

Testing over the real image employing our defect detection and classification approach is to show the performance of our detection and classification method in whole. The steps of our approach are listed in the bottom row of Fig. 1. First, a median filter is adopted to remove the noise while keeping the edges. Second, the morphology operator “Open” is used to estimate the background with non-uniform illumination. A diamond with radius 6 is used for the morphology operation. Furthermore, the estimated background is subtracted. The remainder image has uniform background and the defects are easy to be separated by thresholding. After the image has been binarized, any shape that has less than 8 pixels has been excluded for furthermore processing. Moreover, the image has been segmented, and each separated defect has been labeled. The weight center and a rectangle are then calculated for each

labeled patch. Since the defect shape is variable in size, in order to match the training template, each labeled shape has been resized to 20×20 with weight center as the center of the defect patch. Finally, features are the binary vectors of the pixels values, the defect patches are finally recognized by the trained SVM.

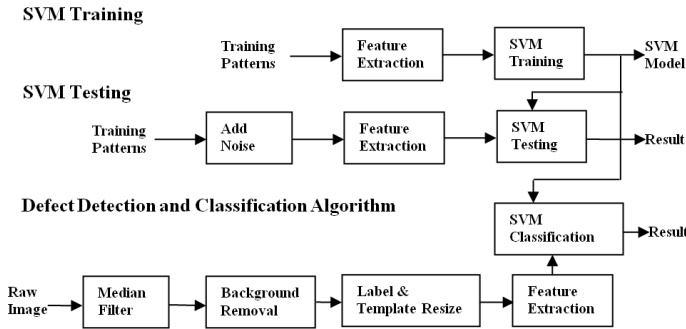


Fig. 1. The flow graph of the proposed automatic defects identification and classification approach

3 Generation of Simulated Data

Rings, semi-circles, clusters, and scratches are the most frequently observed patterns in PCBs and wafers. Samples of them are simulated by the procedure modified from what was proposed by DeNicolao et al. [13], which was also used by Jeong et al. [6] and Palma et al.[14]. With the assumption that electrical failure of a single die occurs independently of the failures of other dies with a probability $p(x, y)$, where x and y are the coordinates of the center of the die.

Since the defect classification is a template matching problem in essence, the procedure has been modified to generate the patterns in 20×20 template images. The value of each pixel is set to be one or zero. All the pixels are zeros if they are outside of the circle of radius 10, and center $(x_c, y_c) = (10, 10)$. Thus, the templates are binary images.

The expressions of the spatial probability $p(x, y)$ corresponding to some common patterns are reported below.

Cluster:
$$p(x, y) = e^{-[(x-x_c)^2/a^2 + (y-y_c)^2/b^2]}$$
,

where the cluster is centered in $(x_c, y_c) = (10, 10)$ and its shape is controlled by the parameters a and b .

Ring:
$$p(x, y) = 1 - e^{-[(x-x_c)^2 + (y-y_c)^2]/2\sigma^2}$$

The ring is centered in $(x_c, y_c) = (10, 10)$ and its width is controlled by the parameter σ .

Semi-circle: By a proper choice of the center (x_c, y_c) , it is possible to shift the ring so as to obtain semi-ring shapes.

Scratch: $p(x, y) = [1 + \sin(2\pi y / 20T + f)] / 2$

This pattern is characterized by repetitive rows of failures (horizontal), whose positions are controlled by the parameters T and f . Here it is adjusted to simulate the scratch or line defect patterns.

For the clusters, the parameters are chosen: $a = 5, 6, 7, 8$, and $b = 10$. Each shape is then rotated by 30 degrees repeatedly, 48 clusters in total are generated. Since the ring has more obvious features than others, there are 13 rings with the deviation range from 10, 10.5, ..., 16. All the semi-circles have the deviation 20; the centers are all the pixels in the circle area where the distance to the center $(x_c, y_c) = (10, 10)$ is 5 with width 1.2. There are 76 semi-circles. The weighted center is found, and the whole template has been shift from weighted center to the center $(x_c, y_c) = (10, 10)$. The pixels outside of the 20×20 have been truncated. The scratch pattern is generated from the repetitive (horizontal) pattern. With $T = 0.8$, $f = 40$, x value of the scratch ranges from 8 to 14. The second set has $T = 0.2$, $f = 10$ with the same x value. Each one has been rotated for every 20 degrees, thus in total, there are 36 generated scratches.

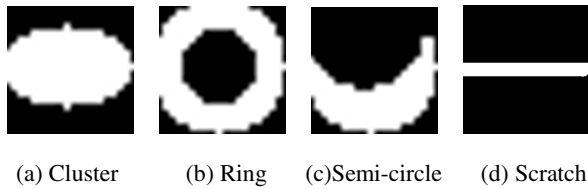


Fig. 2. Simulated defect patterns for SVM training

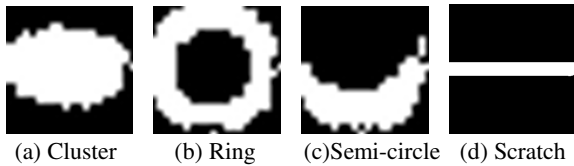


Fig. 3. Simulated defect patterns with noise for SVM testing

Fig. 2 shows the simulated defect patterns for the SVM training. From left to right, they represent the clusters, ring, semi-circle, and scratch respectively. With the simulated data, we avoid the problem that many of the conventional methods have, i.e., the same patterns with different scales; locations or rotations are often classified as different patterns by those methods. On the other hand, our approach is scale/location/rotation invariant. In order to evaluate the classification performance of the SVM algorithms, similarly, another 10 sets of test data are simulated with 10 percent random

noise added. Around the edge, canny filter is used and the pixels are randomized to be 0 or 1 with probability 0.5. Fig. 3 lists the noisy defect patterns for SVM testing.

4 Experiments

In this section, SVM are trained by the simulated training data, and are tested by the simulated test data. And then, the real noisy data is tested using the proposed approach with SVM classifiers.

4.1 Experiments over Simulated Data

Each 20×20 defect pattern in the simulated training data in section 4 is re-arranged as a feature vector. Similarly, the 10 sets of randomized noisy test data are re-arranged. Both training data and test data have 173 samples.

The selection of the parameters of multi-class SVMs is achieved by grid searching, which shows that the linear kernel with $c=0.03$ gets the best classification results. The testing data are tested by the trained multi-classes SVM. All defect patterns are correctly classified. This concludes that SVM can achieve satisfactory results in defect patterns recognition.

The testing results of SVM with are shown in the Table 1. Experiments in this subsection show that can achieve satisfactory results over the simulated data.

Table 1. Simulation results with SVM over simulated test data

Category	Total samples	SVM
Ring	13	13
Semi-circle	76	76
Cluster	48	48
Scratch	36	36

4.2 Experiments over Real Data

For the real noisy image, the proposed approach has been applied to automatically detect the defects and classify the patterns using the SVM. Fig. 4.a is the raw image which is taken from the microscope. The intensities of the image are not uniform. It has been contaminated by environmental light. The lights are from the top, and as a consequence, the pixels in the upper part of the image are brighter than those elsewhere. The pixels outside of the focusing circle are darker than those inside. The defects' sizes range from roughly 10 pixels to more than 100 pixels. The task is to detect all the defects and then classify the rings and semi-circles. Directly thresholding fails to work in such noisy image. Circular Hough transform can detect circle easily. However, rings, semi-circles, and clusters all have bright peaks in the mapping space. Thus it's hard to recognize them separately.

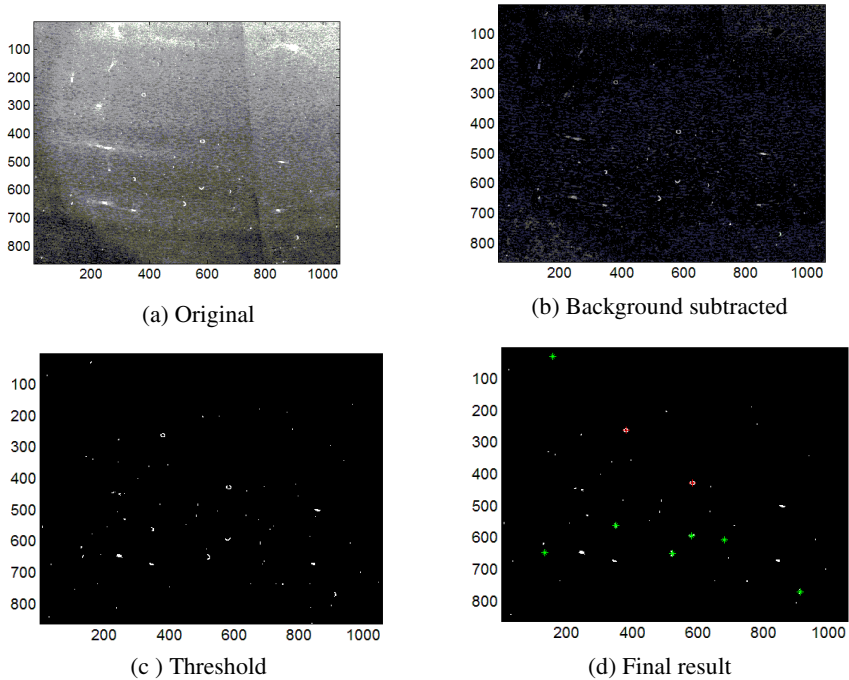


Fig. 4. Results of applying the proposed approach to real data. (a) Original image. (b) Result of applying median filter and background subtraction. (c) Result of thresholding from (b). (d) The final results of the new approach. The detected circles are marked with red “+” sign, while the detect semi-circle are represented by green “*”.

Satisfactory results can be achieved by applying our new approach. The 3×3 median filter is used to remove the noise and small particles, such as pepper and salt. The diamond with radius 6 is used for morphology operation “Open” to estimate the background. Fig. 4(b) shows the result after subtracting the estimated background from the image with non-uniform illumination. The background in the new image is more uniform compared to Fig. 4(a). Thresholding separated all the defects out. The binary image of Fig. 4(c) shows the detected defects. Segmentation helped to find out all disconnected defects. All the defects with more than 8 pixels were labeled. In order to increase the speed, the rectangle boundary and weighted center were calculated for each labeled defect. Since the defect patterns are various in size, each labeled defect has been resized to 20×20 . For semi-circles, the image patch has been shifted when it is resized to 20×20 such that the weight center of the original patch is the center of the new 20×20 defect patch. When preparing for the simulated training data, the defect patterns are rotated, thus it overcomes the shortcoming of SOM and ART1, which misclassify the defect patterns with same shape but different sizes, or locations, or angles as different patterns. Finally, all the modified defects were tested by the trained SVM. In Fig. 4(d) we mark the two rings with red “+” sign, while the detected 7 semi-circles are represented by green “*”.

5 Conclusions and Discussions

Rings, semi-circles, clusters, and scratches are most frequently observed defect patterns both in PCBs and wafers. Each type of patterns associates to some specific cause of the process malfunction. Unsupervised learning algorithms such as SOM and ART1 misclassify some patterns as different defect when they have the same shape but different sizes, different locations, or different angles. Furthermore, when the images have noise and the background is not uniform, most conventional methods failed to achieve good result.

In this paper, a novel approach of automatic identification of defect patterns has been proposed to address these problems. The design of the simulation data, and the defect detecting stage succeed to identify two scale-variant or shift-variant or rotation-variant defect patterns, which in fact belong to the same failure causes. Multi-class support vector machines are employed to classify the defect patterns. The experimental results of applying the new method over real noisy data confirmed the effectiveness of the proposed approach. This new approach can also be easily extended to include other defect patterns, which is in our plan for future work. Comparison with ART1 and other existing algorithms will be our future work.

References

1. Chou, P.B., Rao, A.R., Struzenbecker, C., Wu, F.Y., Brecher, V.H.: Automatic defect classification for semiconductor manufacturing. *Machine Vision and Applications* 9(4), 201–214 (1997)
2. Cunningam, S.P., MacKinnon, S.: Statistical methods for visual defect methodology. *IEEE Transaction on Semiconductor Manufacturing* 11, 48–53 (1998)
3. Chien, C.F., Wang, W.C., Cheng, J.C.: Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications* 33(1), 192–198 (2007)
4. Hsieh, H.W., Chen, F.L.: Recognition of defect spatial patterns in semiconductor fabrication. *International Journal of Production Research* 42(19), 4153–4172 (2004)
5. Jeong, Y., Kim, S., Jeong, M.: Automatic Identification of Defect Patterns in Semiconductor Wafer Maps Using Spatial Correlogram and Dynamic Time Warping. *IEEE Transactions on Semiconductor Manufacturing* 21(4), 625–637 (2008)
6. Gu, N., Cao, Z., Xie, L., Tan, M., Nahavandi, S.: Identification of Concurrent Control Chart Patterns with Singular Spectrum Analysis and Learning Vector Quantization Network. *Journal of Intelligent Manufacturing* (2012), doi:10.1007/s10845-012-0659-0
7. Hu, J., Chen, J., Sundararaman, S., Chandrashekhara, K.: Finite Element Analysis of V-Ribbed Belt/Pulley System with Pulley Misalignment using a Neural Network based Material Model. *Neural Computing and Applications* 18(8), 927–938
8. Su, C.T., Yang, T., Ke, C.M.: A neural-network for semiconductor wafer post-sawing inspection. *IEEE Transactions on Semiconductor Manufacturing* 15, 260–266 (2002)
9. Huang, C.J.: Clustered defect detection of high quality chips using self-supervised multi-layer perceptron. *Expert Systems with Applications* 33, 996–1003 (2007)

10. Chen, F.L., Liu, S.F.: A neural-network approach to recognize defect spatial pattern in semiconductor fabrication. *IEEE Transaction on Semiconductor Manufacturing* 13(3), 366–373 (2000)
11. Lee, J.H., Yu, S.J., Park, S.C.: A new intelligent SOFM-based sampling plan for advanced process control. *Expert Systems with Applications* 20, 133–151 (2001)
12. Liu, S.F., Chen, F.L., Lu, W.B.: Wafer bin map recognition using a neural network approach. *International Journal of Production Research* 40(10), 2207–2223 (2002)
13. DeNicolao, G., Pasquinetti, E., Miraglia, G., Piccinini, F.: Unsupervised spatial pattern classification of electrical failures in semiconductor manufacturing. In: *Workshop on Artificial Neural Networks in Pattern Recognition*, pp. 125–131 (2003)
14. Palma, F., DeNicolao, G., Miraglia, G., Pasquinetti, E., Piccinini, F.: Unsupervised spatial pattern classification of electrical-wafersorting maps in semiconductor manufacturing. *Pattern Recognition Letters* 26(12), 1857–1865 (2005)
15. Sun, J., Zhang, W., Ning, D.: Fault-diagnosis of Flight Control System Based on Neural Network. *Measurement & Control Technology* (5) (2008)
16. Wang, C.H.: Separation of composite defect patterns on wafer bin map using support vector clustering. *Expert Systems with Applications* (2), 2554–2561 (2009)
17. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2), 955–974 (1998)
18. Xie, L., Li, D., Simske, S.J.: Feature dimensionality reduction for example-based image superresolution. *Journal of Pattern Recognition Research* 2, 130–139 (2011)
19. Song, H., Choi, K.K., Lee, I., Zhao, L., Lamb, D.: Adaptive Virtual Support Vector Machine for the Reliability Analysis of High-Dimensional Problems. *Structural and Multidisciplinary Optimization* 47(4), 479–491 (2013)
20. Xie, L., Gu, N., Li, D., Cao, Z., Tan, M., Nahavandi, S.: Concurrent Control Chart Patterns Recognition with Singular Spectrum Analysis and Support Vector Machine. *Computers & Industry Engineering* 64(1), 280–289 (2013)
21. Li, T., Huang, C.: Defect spatial pattern recognition using a hybrid SOM–SVM approach in semiconductor manufacturing. *Expert Systems with Applications* 36(1), 374–385 (2009)
22. Kohonen, T.: The self-organizing map. *Proceedings of the IEEE* 78(9), 1464–1480
23. Wang, Q., Cao, C., Li, M., Zu, H.: A New Model Based on Grey Theory and Neural Network Algorithm for Evaluation of AIDS Clinical Trial. *Advances in Computational Mathematics and its Applications* 2(3), 292–297 (2013)
24. Qian, H., Cao, C., Zu, H., Wang, Q., Li, M.: Exploring the Network Scale-out in virtualized Servers. In: *Proceeding of International Conference on Soft Computing and Software Engineering (SCSE 2013)* (2013)

Illumination Invariant Face Recognition

Guangyi Chen¹, Sridhar Krishnan², Yongjia Zhao³, and Wenfang Xie⁴

¹ Department of Computer Science and Software Engineering, Concordia University,
Montreal, Quebec, Canada H3G 1M8
guang_c@cse.concordia.ca

² Department of Computer and Electrical Engineering, Ryerson University,
Toronto, Ontario, Canada
Krishnan@ee.ryerson.ca

³ State Key Lab. of Virtual Reality Technology and Systems, Beihang University,
ZipCode 100191, No 37, Xueyuan Rd., Haidian District, Beijing, P.R. China
activezyj@126.com

⁴ Department of Mechanical & Industrial Engineering, Concordia University,
Montreal, Quebec, Canada H3G 1M8
wfxie@encs.concordia.ca

Abstract. We present a novel method for face recognition by enhancing the quality of the input face images, which may be too dark due to different lighting conditions. We propose to extract the FFT features or the dual-tree complex wavelet (DTCWT)-FFT features from the enhanced face images and use the Support Vector Machine as a classifier. Our experiments show that our proposed methods compare favourably to the FFT features without image enhancement, and the methods in [1] and [10] for the Extended Yale Face Database B and the CMU-PIE face database.

Keywords: Face recognition, pattern recognition, fast Fourier transform (FFT), sparse FFT, support vector machines (SVM).

1 Introduction

Face recognition is a very important topic in many real-life applications. We briefly review a few existing face recognition methods here. Du and Ward [1] developed an adaptive region-based image enhancement method for face recognition. However, this method introduces defects along the boundary of different image regions. In addition, this method performs histogram on the evenly illuminated face images, which is undesirable. Ruiz-Pinales et al. [2] also proposed a shift invariant support vector machines face recognition system.

In this paper, we present a new method for face recognition by using the FFT or dual-tree complex wavelet (DTCWT)-FFT features [4] and Support Vector Machines (SVM) as a classifier. We enhance the visual quality of the face images by using histogram equalization, and blend the resulting image with the original face image. We then extract the FFT or DTCWT-FFT features from the enhanced images. We feed the extracted features to the SVM in order to classify the input face to one of the faces in the face database. Our recognition rates are higher than the FFT features without image

enhancement, and the methods in [1] and [10] for the Extended Yale Face Database B and they are comparable for the CMU-PIE illumination face database.

2 Illumination Invariant Face Recognition

In this paper, we present a novel method for face recognition. Due to difference in illumination conditions, the captured face images may be very dark. In order to get better recognition rates, we enhance the visual quality of the face images by using histogram equalization. We blend the histogram-equalized image with the original face image. We also calculate the mean of the original face image and compare the mean with a threshold ($t=50$). If the mean is less than the threshold, then we set $\alpha=2/3$, and $\alpha=1/2$ otherwise. We use this α to blend the histogram-equalized image and the original face image. We extract the FFT or the DTCWT-FFT features from the enhanced face images and then use the Support Vector Machines (SVM) to classify the face images into one of the images in the face database.

Our proposed algorithms in this paper can be summarized as follows:

- (a) Calculate the mean of the face image f , denote it as \bar{f} .
- (b) Perform histogram equalization (denoted as Matlab command 'histeq') on the face image f :

$$g = \text{histeq}(f).$$

- (c) If $\bar{f} < t$ ($t=50$ is a threshold), $\alpha=3/4$; and $\alpha=1/2$ otherwise.

- (d) Blend the face image f with the histogram equalized image g :

$$h = \alpha \times g + (1 - \alpha) \times f$$

- (e) Extract the FFT or the DTCWT-FFT features from the preprocessed face image and take the magnitudes:

$$|FFT(u, v)| = \frac{1}{MN} \sum_{m=1}^M \sum_{n=1}^N h(m, n) e^{-i(um/M + vn/N)}.$$

We can replace the FFT in this step by sparse FFT ([8], [9]), which is much faster than the standard FFT when the face image is sparse in the frequency Domain.

- (f) Classify this face image into one of the existing faces in the database by using the extracted features and SVM as a classifier.

The computational complexity of our proposed method can be estimated as follows. Let the input face images have M rows and N columns. Also, let the number of training images be P and the number of testing images be Q . Then the computational complexity of our proposed method is $O(PMN \log(MN))$ for training, and $O(QMN \log(MN))$ for testing. If we replace the FFT in step (e) of our algorithm with sparse FFT ([8], [9]), then the computational complexity is $O(PK \log(MN))$ for training, and $O(QK \log(MN))$ for testing, where K is the number of nonzero sparse Fourier coefficients.

3 Experimental Results

We conducted experiments for face recognition by using the Extended Yale Face Database B [11]. The database has 2414 frontal face images of 38 individuals.

The cropped and normalized 192×168 face images were captured under various laboratory - controlled lighting conditions. Fig. 1 shows a sample of the face images in this database under different lighting conditions and Fig. 2 shows the enhanced face images by our proposed method in this paper. It is easy to see that our image enhancement method can do an excellent job for invariant face recognition.



Fig. 1. A sample of the face images under different lighting condition of the Extended Yale-B face database



Fig. 2. A sample of the enhanced face images under different lighting condition of the Extended Yale-B face database



Fig. 3. A sample of the face images under different lighting condition of the CMU-PIE illumination face database



Fig. 4. A sample of the enhanced face images under different lighting condition of the CMU-PIE illumination face database

We also tested the CMU-PIE illumination face database [12], which contains 68 individuals with 21 illumination variations. Fig. 3 shows a sample of the face images in this database under different lighting conditions whereas Fig. 4 shows the enhanced face images for this database. We chose half of the face images for training and the other half for testing. We apply our proposed method to this database and extract the FFT or the DTCWT-FFT features, and then choose the SVM as a classifier. Cross-validation is utilized to select the best parameter γ in the radial basis function (RBF) of the SVM kernel:

$$k(x, y) = e^{-\gamma\|x-y\|^2}.$$

We select to use the LIBSVM [3] for our experiments. We compare our proposed methods, namely the Enhanced FFT (E-FFT) and the DTCWT-FFT methods with the FFT without enhancement in Table 1. It can be seen that our proposed methods yield higher classification rates but with less number of features. In Table 1, the DTCWT-FFT 1 means that we use features in LH, HL, and HH for scales 2 and 3. The DTCWT-FFT 2 means that we use the sum of features in LH, HL, and HH for scales 2 and 3, respectively. Even though the DTCWT has approximate shift invariant property and it has been successfully applied to character recognition, shape recognition, and signal denoising ([5], [6], [7]), it is not as good as our E-FFT in this paper for face recognition for the Extended Yale Face Database B. However, all our proposed three methods perform equally well for the CMU-PIE illumination face database.

Table 1. A comparison of the FFT without image enhancement, the DTCWT-FFT 1 (LH, HL, HH for scales 2 and 3), the DTCWT-FFT 2 (sum of LH, HL, HH for scales 2 and 3, respectively), and the method E-FFT. Our proposed methods use less number of features, but with higher recognition rates for the Extended Yale Face Database B and the CMU-PIE face database. The best results are highlighted in bold font.

		Dimension	Recognition rates (%)	Dimension	Recognition rates (%)
Extended Yale-B	FFT	169	82.57	400	81.74
	DTCWT-FFT 1	150	88.90	600	91.45
	DTCWT-FFT 2	168	76.15	432	86.02
	E-FFT	169	93.17	400	94.98
CMU-PIE	FFT	169	99.56	400	99.71
	DTCWT-FFT 1	150	99.71	600	99.71
	DTCWT-FFT 2	168	99.71	450	99.71
	E-FFT	169	99.71	400	99.71

We compare our proposed method E-FFT with two existing methods [1] and [10]. For the Extended Yale-B database, the method [1] obtained an average recognition rate of $(100+84.4+90.4+80.5)/4=88.83\%$ for this database. Our proposed method achieved 94.98% recognition rate with 400 Fourier coefficients, which is much higher than the recognition results reported in [1]. Xie et al. [10] proposed an illumination normalization technique on large and small scale features for robust face recognition. The average recognition rate of [10] is $(100+87.1+87.6+84.8)/4=89.88\%$ for this database. This recognition rate is higher than the recognition rate of [1]. However, it is

not as good as our proposed method. This demonstrates the success of our proposed method for robust face recognition.

For the CMU-PIE illumination database, both [1] and [10] achieved 100% correct classification rate whereas our proposed methods obtain 99.71% correct classification rates. From Fig. 13 of [1], we know that the method [1] uses cropped face images, which is different from our face images used in this paper. We believe that our proposed methods are good choices for illumination-robust face recognition.

4 Conclusions

In this paper, we have proposed a new method for face recognition by enhancing the face images and then extracting the FFT or the DTCWT-FFT features. SVM is used as a classifier and the best parameter in the SVM kernel is selected by cross-validation. Our proposed methods are better than the FFT without image enhancement, and the methods in [1] and [10] in terms of correct recognition rates for the Extended Yale Face Database B. Our methods are comparable to the methods [1] and [10] for the CMU-PIE illumination face database. Compared with existing methods for face recognition, our proposed method in this paper is more suitable for recognizing face images with uneven illumination. Our method is very simple to implement and it achieves state-of-the-art recognition rates for the Extended Yale Face Database B and the CMU-PIE illumination face database.

Acknowledgments. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Beijing Municipal Science and Technology Plan: Z111100074811001.

References

1. Du, S., Ward, R.K.: Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions. *IEEE Transactions on Circuits and Systems for Video Technology* 20(9), 1165–1175 (2010)
2. Ruiz-Pinales, J., Acosta-Reyes, J.J., Salazar-Garibay, A., Jaime-Rivas, R.: Shift invariant support vector machines face recognition system. *World Academy of Science, Engineering and Technology* 16, 947–951 (2008)
3. Chang, C.C.: LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3), Article no. 27 (2011)
4. Kingsbury, N.G.: Complex wavelets for shift invariant analysis and filtering of signals. *Journal of Applied and Computational Harmonic Analysis* 10(3), 234–253 (2001)
5. Chen, G.Y., Bui, T.D., Krzyzak, A.: Invariant pattern recognition using radon, dual-tree complex wavelet and Fourier transforms. *Pattern Recognition* 42(9), 2013–2019 (2009)
6. Chen, G.Y., Zhu, W.P.: Signal denoising using neighbouring dual-tree complex wavelet coefficients. *IET Signal Processing* 6(2), 143–147 (2012)
7. Chen, G.Y., Xie, W.F.: Contour-based feature extraction using dual-tree complex wavelets. *International Journal of Pattern Recognition and Artificial Intelligence* 21(7), 1233–1245 (2007)

8. Hassanieh, H., Indyk, P., Katabi, D., Price, E.: Simple and practical algorithm for sparse Fourier transform. In: SODA (2012)
9. Hassanieh, H., Indyk, P.: Katabi and Price, E.: Nearly optimal sparse Fourier transform. In: STOC (2012)
10. Xie, X., Zheng, W., Lai, L., Yuen, P.C.: Face illumination normalization on large and small scale features. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
11. Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(5), 684–698 (2005)
12. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(2), 1615–1618 (2003)

Evaluation of Sampling Methods for Learning from Imbalanced Data

Garima Goel¹, Liam Maguire¹, Yuhua Li¹, and Sean McLoone²

¹ School of Computing and Intelligent Systems, University of Ulster, UK
{lp.maguire,y.li}@ulster.ac.uk

² Department of Electronic Engineering, National University of Ireland Maynooth, Ireland
semcloone@eeng.nuim.ie

Abstract. The problem of learning from imbalanced data is of critical importance in a large number of application domains and can be a bottleneck in the performance of various conventional learning methods that assume the data distribution to be balanced. The class imbalance problem corresponds to dealing with the situation where one class massively outnumbers the other. The imbalance between majority and minority would lead machine learning to be biased and produce unreliable outcomes if the imbalanced data is used directly. There has been increasing interest in this research area and a number of algorithms have been developed. However, independent evaluation of the algorithms is limited. This paper aims at evaluating the performance of five representative data sampling methods namely SMOTE, ADASYN, BorderlineSMOTE, SMOTE-Tomek and RUSBoost that deal with class imbalance problems. A comparative study is conducted and the performance of each method is critically analysed in terms of assessment metrics.

Keywords: imbalanced data; sampling methods, support vector machines.

1 Introduction

The problem of imbalance learning refers to the problem of class imbalance among the datasets where one class contains much more examples than the other class thus introducing the concept of the majority and the minority classes. The above situation creates a challenge for classifiers that focus on the overall accuracy optimization without bothering about the relative data distribution of each class. Hence, the classifier is biased and driven towards the accurate classification of the class containing many data examples or the majority class at the same time overlooking the class containing less number of examples or the minority class [5]. Unfortunately, the problem is quite prominent in the concept learning domains as most of them suffer from the problem of class imbalance. The examples of such imbalances can be seen in some diverse application such as: production line monitoring, where the majority of the products are non-defective and only small minority are defective, fraudulent phone calls detection, fault monitoring for gearbox of the helicopter, oil spills detection, etc.

However the limited number of examples in minority class is usually more important and good performance on the minority is desirable.

Machine learning relies on models built on data. Most learning algorithms implicitly assume that the data distribution is balanced among different classes. If the data distribution exhibits significant imbalance among classes, learning would be biased towards the majority class, resulting in unacceptable errors on the minority class. Accuracy on the minority class is often the ultimate goal of many applications such as medical diagnosis and capital markets, where there is sufficient data for normal trading patterns, but very little data about irregular trading patterns. An efficient strategy for achieving reliable learning from imbalanced data is to employ a sampling scheme to balance the data. It could generate samples for the minority class to bring up the size of the minority class, or re-sample the majority class to reduce the size of the majority class. As the size of the minority class is often very small for real applications, re-sampling the majority class to balance data would result in small overall data size. Therefore sample generation for the minority class is preferred for many applications. However the study of the performance and the applicability of individual sampling methods is limited.

This paper aims at evaluating five different sampling methods namely, SMOTE [3], ADASYN [8], BorderlineSMOTE [6], SMOTETomek [1] and RUSBoost [14] using eight different datasets on the basis of assessment metrics including overall accuracy, recall, precision, F-measure and G-mean. These methods have been assessed using the classifier of Support Vector Machine (SVM). The performance of each method on benchmark datasets under each assessment criterion is analysed.

This paper is organised as follows: Section 2 establishes the background with some of the popular methods or techniques used in solution to the problem of imbalanced learning. Section 3 explains the selection of the five sampling methods and describes the experimental procedure including the dataset analysis, evaluation metrics and the experimental results and analysis. Section 4 concludes the paper with a brief description of the achieved results.

2 Background

When imbalanced data is passed onto a standard algorithm, the learning rules are biased towards the majority class data as the minority class data are usually fewer in number. This is because the learning rule minimizes the error (or maximise accuracy) on the training data and the minority examples in an imbalanced data are often under-represented. Many algorithms have been developed to handle the learning problems of imbalanced data, they can be classified into four categories: sampling-, cost-, kernel- and active learning-based methods. For a comprehensive review, readers are referred to He et al. [7].

In this study, sampling methods are taken as the main focus and a brief overview of these techniques are provided. The methods of sampling concentrate on balancing the distribution of the samples belonging to the minority and the majority class in order to deal with the problem of imbalanced data, though it is still not clear what the true

class distribution should be to effectively handle the imbalanced problem [4]. Studies show that balanced data help the classifiers learn more accurately than imbalanced data [15], [11]. The literature presents two major sampling techniques namely, undersampling and oversampling.

2.1 Sampling Strategies

Random Oversampling and Undersampling: Random oversampling is a method in which examples from the minority class are randomly selected and are augmented to the original data in order to increase the proportion of the minority samples in the dataset. Whereas, in case of undersampling some of the randomly selected examples from the majority class are removed from the dataset in order to equalise the presence of both minority and majority samples.

Informed Undersampling: Some of the methods under this category includes EasyEnsemble, BalanceCascade [12], K-nearest neighbour (KNN) classifier method and one-sided selection (OSS) [10]. The methods aim at reducing information loss that occurs due to the random undersampling. The theme of these algorithms is to develop a learning system that actually helps in the selection of those majority samples whose removal will lead to least information loss.

Synthetic data generation: Synthetic minority oversampling technique (SMOTE) [3] is a powerful approach for generating artificial data based on the similarities between the feature space of the existing minority class samples. The algorithm generates some random points in between the two specified vector points and thus more generalising the minority class decision region.

Adaptive synthetic sampling: This approach includes methods like BorderlineSMOTE [6] and ADASYN [8]. The objective of this approach is to overcome the drawbacks of the synthetic sampling approach in terms of variance and over generalization. The synthetic samples are generated adaptively in consideration with the neighbouring examples so as to avoid overlapping. The ADASYN algorithm adaptively decides the number of synthetic samples that needs to be generated for the minority class samples according to their difficulty level of learning. The aim of the algorithm is two-fold: adaptive learning and bias reduction. Whereas, BorderlineSMOTE algorithm focuses on over-sampling and strengthening of only those minority class samples that lie on the borderline or nearby areas of a class. According to the researchers, borderline examples play the most important role in a classification because it is these elements that a classifier needs to learn as appropriately as possible to give the best prediction results.

Sampling using data cleansing technique: The objective of this approach is to reduce the between class unwanted overlapping that is introduced upon the application of sampling algorithm. Some of the methods include one-sided selection (OSS), integration of condensed nearest neighbour rule and Tomek links (CNN+Tomek Links) [1], neighbourhood cleansing rule (NCL) [11], integration of SMOTE and edited nearest neighbour (SMOTE+ENN) and integration of SMOTE and Tomek links

(SMOTE+Tomek). SMOTETomek [1] is a technique of data cleansing that involves the removal of the Tomek links which are the pairs of examples that are in closest proximity to each other or called the nearest neighbours but belongs to opposite classes. The resulting dataset is free from the between class crowded overlaps so that all nearest neighbours placed at the minimal distance belongs to one class only.

Cluster-based sampling technique: Cluster-based oversampling algorithm deals with the problem of within-class imbalance and between-class imbalance in tandem based on the imbalance ratio among the clusters. One of the algorithms is CBO (Cluster Based Over-sampling) [9].

Sampling and boosting integration: Some of the methods include SMOTEBoost [2], DataBoost-IM, JOUSBoost [7], RUSBoost. The aim of these approaches is to combine sampling strategies with the ensemble learning techniques so that the classifier can learn more at each iteration to provide a better performance. RUSBoost [14] is a technique of data sampling that requires no intelligent effort to remove samples from a training set and the samples are removed until the desired distribution of the class is achieved. It is an approach that makes use of the random under-sampling technique along with the boosting technique ADABOOST.M1. The random undersampling in RUSBoost is performed in order to balance the distribution while the ADABOOST [13] focuses on the improvement on the classifiers performance.

3 Experiments

This study evaluates the performance of five representative sampling methods for learning from imbalanced data based on assessment metrics. The five methods are: SMOTE, SMOTETomek, BorderlineSMOTE, ADASYN, and RUSBoost. All these methods are chosen from the category of sampling strategies in order to range them from the basic and popular SMOTE to the improved approaches and RUSBoost which is the latest combination of two classic strategies, random undersampling and AdaBoost. Also, the selected range of methods covers almost all the subcategories of sampling strategies. The first four methods namely, SMOTE, SMOTETomek, BorderlineSMOTE and ADASYN are consecutive improved versions of the oversampling technique whereas RUSBoost involves a technique of undersampling for its sampling strategy in the field of imbalanced learning. The base learner benchmark used in this study is the support vector machine as it is considered to be a well recognised state-of-the art classifier.

In this study the performance of the five data sampling methods are compared and examined using five different performance metrics that focus on handling the imbalanced class data. These sampling techniques are applied before building the SVM models. Results based on the performance of the metrics are calculated using SVM as a base classifier and each experiment is repeated 10 times in order to avoid any bias. The results obtained are the averaged results of the repetitions. The training and test dataset were divided in the ratio 50:50 to perform grid search.

3.1 Datasets

The experiments were conducted on eight different datasets from the UCI Repository. These datasets have variations in their size as well as in their class proportions. The experiment deals with the binary classification of data and therefore the datasets are divided into two subsets of majority and minority. Majority class data are identified with the negative labels whereas the positive labels are provided to the minority class data. Table 1 describes the datasets in terms of their size, number of attributes (dimensions) and the percentage of distribution of minority class data in the original dataset and the datasets created after application of sampling methods. The size of the dataset is the size after removing all the NaN (not a number) values from the original dataset.

Table 1. Minority size distribution (%) for datasets *

Data Name	Size	Dim	Original	SMOTE ADS	BDS	STM	RUB
Hepatitis	80	3	16	49	46	43	49
Haberman	306	3	26	42	46	47	39
BreastCancerO	683	10	35	50	43	60	50
Australian	690	14	32	49	41	60	48
Pima	768	8	35	50	44	53	50
Abalone	4177	29	06	50	49	15	50
Spambase	4601	57	39	50	45	51	50
Sensor-readings	5456	24	40	57	47	47	57

*ADS – ADASYN, BDS – Borderline SMOTE, STM – SMOTE Tomek; RUB – RUSBoost

3.2 Evaluation Metrics

Let {p, n} be the positive and negative labels for the two-classes respectively and {P, N} be the positive and negative labels that are predicted. A confusion matrix is readily developed using the above parameters for assessing the performance of the classification for imbalanced learning. The confusion matrix evaluates the values of the true and false positives and the negatives as determined by the classifier as shown in (Fig 1).

	p	n
P	TP (True Positive)	FP (False Positive)
N	FN (False Negative)	TN (True Negative)

Fig. 1. Confusion matrix

Some of the traditional evaluation metrics that are used for the purpose of accessing performance are the overall accuracy and error rate. The accuracy and the error rate can be defined using the following formula:

$$Accuracy = \frac{TP + TN}{P_c + N_c}, \quad ErrorRate = 1 - Accuracy .$$

where, $P_c = TP+FN$ and $N_c = FP+TN$. In the paper in addition to overall accuracy and the error rate, a few standard evaluation criteria including precision, recall, F-measure and G-mean are utilised for the assessment of the classifier's performance [7]. The formulas for the above measures are as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN},$$

$$F - measure = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Recall + Precision}$$

here, β is the coefficient that is used to set the importance of Precision versus Recall relatively and is usually equal to 1.

$$G - mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}}$$

Precision can be understood as a measure of exactness as it determines the number of positive samples that are correctly labelled among all the positive labels of a class, whereas recall determines the measure of completeness by identifying the number of correctly labelled positive samples among the predicted positive labels [7]. F-measure is another effective measure for evaluating a classifier's performance. It is a combined factor of both precision and recall along with the weighted importance as determined by the parameter β , a coefficient that is adjusted by a user. G-mean is a degree measure of inductive bias which is evaluated as the ratio of positive accuracy over negative one.

3.3 Experimental Results

All five methods are implemented in Matlab. The datasets are sampled using these algorithms separately and the resulting re-sampled data is then passed on to a classifier for evaluation purposes. The kernel function used in the SVM classifier is a radial basis function. Some of the other parameters set to perform the sampling processes includes, $K=5$ which means that the number of nearest neighbours equals five for SMOTE, ADASYN, BorderlineSMOTE and SMOTETomek which is taken as standard value set as used for SMOTE [3]. ADASYN uses 0.75 as the threshold value as given in [8]. For the RUSBoost algorithm the post sampling dataset is distributed in the varied ratio of 35, 40, 45 and 50 where the values represents the percentage of examples for minority class. This ratio of the majority to minority data sometimes depends on the class distributions in the original dataset. Random undersampling reduces the size of the dataset considerably and thus decreasing the training time of the models. Although a major drawback of information loss occurs in the process of undersampling due to deletion of the randomly selected samples from the training set.

The experimental results are presented in Tables 2 to 4. Table 2 shows the performance results of accuracy, F-measure and G-mean (in the order in each cell of the table, respectively). Whereas, Table 3 shows the performance results of recall and

precision on each dataset (in the order in each cell of the table, respectively). For every algorithm applied to a dataset, the best performance in each category such as accuracy, precision, recall, F-measure and G-mean are highlighted. Table 4 identifies the total winning times of an algorithm in each type of evaluation metrics thus enabling to better summarise the performance of each method under various evaluation criteria of imbalanced learning problems. Based on the overall results of evaluation as shown in Table 2, 3 and 4, the RUSBOOST method outperforms by winning the maximum number of times. Next, in the sequence is SMOTETomek which has also shown a good number of winning times. The BorderlineSMOTE falls third in the winning sequence and has comparative results like SMOTETomek in the total number of winning times. The other methods like SMOTE and ADASYN shows quite a lacking performance in comparison with the other methods.

Table 2. Results of Accuracy, F-measure and G-mean (in the order in each cell of the table, respectively)

Data Set	Original	SMOTE	ADS	BDS	STM	RUB
Breast Cancer	0.9677, 0.9747, 0.9564	0.9730, 0.9749, 0.9734	0.9562, 0.9611, 0.9529	0.9821, 0.9794, 0.9862	0.9672, 0.9649, 0.9687	<u>0.9927</u> , <u>0.9944</u> , <u>0.9897</u>
Haberman	0.6471, 0.7404, 0.5606	0.6684, 0.7130, 0.6490	0.6555, 0.6635, 0.6551	0.6933, 0.6570, 0.6938	<u>0.7574</u> , 0.7773, <u>0.7347</u>	0.7338, <u>0.8038</u> , 0.6614
Hepatitis	0.9000, 0.9429, 0.7521	0.9091, 0.9167, 0.9113	0.8548, 0.8657, 0.8515	0.9455, 0.9552, 0.9325	<u>0.9833</u> , <u>0.9825</u> , <u>0.9826</u>	0.8780, 0.9231, 0.7620
Pima Indians	0.7344, 0.7935, 0.7014	0.7800, 0.7727, 0.7812	0.7478, 0.7601, 0.7461	<u>0.7941</u> , 0.7683, <u>0.7957</u>	0.7597, 0.7511, 0.7601	0.7532, <u>0.8081</u> , 0.6979
Spambase	0.7861, 0.8488, 0.8488	0.8436, 0.8608, 0.8612	0.7677, 0.8212, 0.8137	<u>0.8792</u> , <u>0.8814</u> , <u>0.8813</u>	0.7837, 0.8203, 0.8212	0.8230, 0.8720, 0.8729
Australian	0.6058, 0.7190, 0.5001	0.6447, 0.6233, 0.6504	0.5995, 0.6681, 0.5762	<u>0.6661</u> , 0.5279, 0.6530	0.6598, 0.6275, <u>0.6686</u>	0.6522, <u>0.7474</u> , 0.5784
Abalone	0.8343, 0.9049, 0.4833	0.8764, 0.8747, 0.8765	0.8466, 0.8385, 0.8506	0.8701, <u>0.9193</u> , 0.7229	<u>0.8860</u> , 0.8819, <u>0.8868</u>	0.8263, 0.8992, 0.4868
Sensor reading	0.8699, 0.8960, 0.8756	<u>0.9303</u> , <u>0.9193</u> , <u>0.9289</u>	0.8686, 0.8774, 0.8685	0.8965, 0.9036, 0.8969	0.9226, 0.9074, 0.9224	0.8832, 0.9057, 0.8872

Being specific, Table 4 shows the winning times of all the considered methods for various assessment metrics such as overall accuracy, recall, precision, F-measure and G-mean. As per the table RUSBoost is a good performer and wins the maximum number of times whereas SMOTETomek falls second with a comparative win. A closer analysis shows that though RUSBoost is the highest performer in terms of winning times but the performance is much better in case of Recall and Precision and F-measure where it is a winner for 14 times in total as compared to other metrics like overall accuracy and G-mean where the performance is quite less. This shows

that the method quite effectively deals with the learning of minority class samples as the recall and precision are the two measures that determines the completeness and exactness of the correctly labelled positive class samples. Also, SMOTETomek is a good performer in case of overall accuracy and G-mean which suggests that the SMOTETomek is a method that provides improved accuracy for both the majority and minority class examples and deals with the completeness of both.

This suggests that all these methods are limited or bounded to certain criteria or conditions under which they perform better. Or in other words, there is a need to set certain rules of thumbs or criteria for using each method, so that the method performs well in its specified range of datasets.

Due to the random nature of the sampling methods that are applied to various datasets, the classifier's performance varied each time the algorithm is run which provided a certain variations in the results. Therefore, the results reported in Table 2, 3 and 4 are the average values of 10 times of running in order to obtain reliable results and to avoid any biases. Also, in case of RUSBoost varying the ratio of majority to minority class distribution affected the results each time they were executed. This raises an important point of discussion that whether there is a need of some standardized measure for the distribution of classes and also to what extent does the random nature of a method affect the class distributions.

Table 3. Results of Recall, Precision (in the order in each cell of the table, respectively)

Data Set	Original	SMOTE	ADS	BDS	STM	RUB
Breast cancer	<u>1.0000</u> , 0.9507	0.9907, 0.9596	0.9813, 0.9417	<u>1.0000</u> , 0.9596	0.9880, 0.9429	<u>1.0000</u> , <u>0.9888</u>
Haberman	0.7857, 0.7000	0.6833, 0.7455	0.6827, 0.6455	0.7010, 0.6182	0.7069, <u>0.8632</u>	<u>0.8750</u> , 0.7434
Hepatitis	0.9429, 0.9429	0.8919, 0.9429	0.9063, 0.8286	<u>1.0000</u> , 0.9143	0.9655, <u>1.0000</u>	0.9677, 0.8824
Pima indians	0.8066, 0.7809	0.8026, 0.7450	0.8136, 0.7131	0.8070, 0.7331	0.7682, 0.7348	<u>0.8163</u> , <u>0.8000</u>
Spambase	0.7425, 0.9907	0.7756, 0.9670	0.7136, 0.9670	<u>0.8373</u> , 0.9304	0.7112, 0.9689	0.7776, <u>0.9925</u>
Australian	0.7073, 0.7311	0.6979, 0.5630	0.6639, 0.6723	0.6229, 0.4580	0.7226, 0.5545	<u>0.7396</u> , <u>0.7553</u>
Abalone	0.9757, 0.8437	0.8834, 0.8662	0.8981, 0.7863	0.9786, <u>0.8667</u>	0.9019, 0.8627	<u>0.9863</u> , 0.8263
Sensor reading	0.8567, 0.9392	0.9190, 0.9196	0.8697, 0.8852	0.8928, 0.9147	<u>0.9210</u> , 0.8943	0.8735, <u>0.9403</u>

Table 4. Winning Times of methods over assessment parameters

Methods	Accuracy	Recall	Precision	F_value	G_mean
Original	0	1	0	0	0
SMOTE	1	0	0	1	1
ADASYN	0	0	0	0	0
BorderlineSMOTE	<u>3</u>	3	1	2	2
SMOTETomek	<u>3</u>	1	2	1	<u>4</u>
RUSBoost	1	<u>5</u>	<u>5</u>	<u>4</u>	1

Also, there is a case when the winner among various methods is the original dataset or the performances are nearly comparable which in itself is an issue of discussion as the only purpose of applying a sampling method is to increase the performance of a classifier. So, if the original dataset is good enough to provide the satisfying performance, a question is raised on the underlying principle of the imbalanced learning methods in terms of their extent of providing help with learning abilities. Though there are many current research efforts undergoing in the area of imbalanced learning focussing on the development of the algorithms that provide improved classification accuracy but still the concept lacks and a further work is required for answering such critical questions.

4 Discussion and Conclusion

Imbalanced learning is a very critical and a challenging problem in the field of data engineering and knowledge discovery where the cost of misclassification can be very high. In recent years, tremendous efforts have been invested in the development of sampling methods to balance data and a number of sampling methods have been developed, and each of which has its strengths and weaknesses. However the study of the performance and the applicability of individual sampling methods is limited. This paper focuses on investigating the performance of five such learning algorithms namely, SMOTE, ADASYN, BorderlineSMOTE, SMOTETomek and RUSBoost using eight different datasets on the basis of assessment matrices including overall accuracy, recall, precision, F-measure and G-mean for achieving reliable learning from imbalanced data. These methods have been assessed using support vector machine. The results show that RUSBoost outperforms all the methods by winning the maximum number of times in terms of recall and precision and also the highest overall. Next, in the winning list is SMOTETomek which also has shown good results by winning quite a considerable number of times in terms of overall accuracy and G-mean. BorderlineSMOTE also had its highest number of wins in terms of accuracy. The performance of the other two methods SMOTE and ADASYN was inferior.

The research needs be broadened to set certain benchmarks for different types of algorithm and identifying the underlying conditions under which they can perform better. Also, there are certain critical questions that need to be answered concerning the random nature of the algorithms and what effects do they have on the distribution of the datasets as was the case of RUSBoost in the experiment where the variation in the ratio of majority and minority affected the performance of the classifier. The question also remains unanswered for the cases where the original dataset performs better than any of the methods which indicates that there may be some gap in the understanding and applications of these approaches and whether the balanced distribution is always the correct distribution.

Acknowledgements. The financial support of the Irish Centre for Manufacturing Research and Enterprise Ireland (grant CC/2010/1001) are gratefully acknowledged.

References

1. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 6(1), 20–29 (2004)
2. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: Improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) *PKDD 2003. LNCS (LNAI)*, vol. 2838, pp. 107–119. Springer, Heidelberg (2003)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321–357 (2002)
4. Chawla, N.V., Cieslak, D.A., Hall, L.O., Joshi, A.: Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery* 17(2), 225–252 (2008)
5. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20(1), 18–36 (2004)
6. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: *Advances in Intelligent Computing*, pp. 878–887 (2005)
7. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9), 1263–1284 (2009)
8. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *IJCNN 2008*, pp. 1322–1328 (2008)
9. Kotsiantis, S., Kanellopoulos, D., Pintelas, P.: Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30(1), 25–36 (2006)
10. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: *The 14th International Conference on Machine Learning*, pp. 179–186 (1997)
11. Laurikkala, J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution. In: *AI in Medicine in Europe: Artificial Intelligence Medicine*, pp. 63–66 (2001)
12. 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)
13. Rätsch, G., Onoda, T., Müller, K.R.: Soft margins for AdaBoost. *Machine Learning* 42(3), 287–320 (2001)
14. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 40(1), 185–197 (2010)
15. Weiss, G.M., Provost, F.: The Effect of Class Distribution on Classifier Learning: An Empirical Study. Technical Report ML-TR-43, Dept. of Computer Science, Rutgers Univ (2001)

A Genealogical Study of the Origin of Pashtuns

Hameed Ullah Khan and Nasir Ahmed

University of Engineering and Technology, Peshawar, Pakistan
hameed.leo@gmail.com, n.ahmad@nwfpuet.edu.pk

Abstract. The study of the descent of different races and the investigation of their common ancestors has been of the interest of different disciplines since long. With the advancement of the genetic studies in the recent decades, the researchers have been attracted to utilize the potential of the genealogical studies to find the origin of different races of the world. The origin of Pashtuns is not is one of the most disputed both among historians and pashtuns themselves. This study aims to investigate about the origin of the Pashtuns based on the hypothesis drawn from the most popular claims. The genetic data for the 718 unrelated male subjects was studied. 16 Y-STR binary markers were used in this study. AMOVA was carried out for the data by using the YHRD online service and haplotype search was also carried out using YHRD. The results obtained are compared with the previously conducted similar studies.

Keywords: Pashtuns Origin, Haplotype, Haplogroup, YHRD.

1 Introduction

Pashtuns are one of the major ethnic groups of the world having its population stretched from South to Central Asia, besides a sizable population in the Arabian Gulf, Europe and North America. Despite their vibrant rule in the history of the region, their ethnic origin has always been in shadows. Very scanty data is available about the lineage and genealogy of the Pashtuns. Many oral traditions are present about the origin of the Pashtuns among which the dominant claims are, being descendent of Alexander's soldiers, of Jewish origin and Aryan; however none of these claims has been so far authenticated or rejected clearly. A few early studies about the genealogy of the Pashtuns suffer either from a lack of access to the authentic data or are limited in scope. The study performed for Y-STR profiling for the two afghan populations has the same issue of being limited to the Afghan population only. These studies performed in Pakistan had so far their emphasis on DNA variations among populations and Y-STR profiling. None of the studies referred to had addressed the origin of Pashtuns and none had been carried out collectively for differently located population sizes. In order to find out (search) further about their origin the Y-chromosome of males will be studied to find out the Y-chromosomal Adam or go as far as possible. The non-recombining region of the Y-chromosome will be taken into account in this study as it does not tend to change over the course of time other than unnatural reasons. The Y-STR typing and the haplotype searching pattern for

different Haplogroups approach will be followed. The haplotypes found out using the Y-STR markers will then be used to find out the Haplogroups that they could possibly belong to and look for possible matches from the claimed populations in the database using the YHRD and further on AMOVA using F_{st} (fixation index) values will be calculated to find out any resemblance to previous generations and calculate population pairwise comparison. Further on the study would be conducted in order to dig out more about the origin claims and also the oral traditions would be scrutinized.

2 Related Work

A number of genetic studies of the Pashtuns have been conducted recently by researchers of various universities and research groups. The Greek ancestry of the Pashtuns of Pakistan has been investigated in [1]. In this study, the claim of the three populations of the region, i.e. the Pashtuns, the Kalash and the Burusho, to have descended from the soldiers of Alexander, has been considered. 16 multi allelic Y-STRs were typed and studied to find out the different haplotypes. The YHRD was used as a source to find out the resemblance to any past populations. This research was carried out in 2006 with a limited data and so only 9 Y-STRs were typed and searched for in the YHRD database. The important finding of the study was that Pashtun population contours around Europe and Macedonia. Out of the three populations the Pashtun population was least distant from the Greeks.

In [2], the study was carried out on two Afghan populations i.e. the Pashtuns of the Southern Afghanistan as well as the Northern Afghanistan. The population clustering was done so by taking the Hindu Kush Mountain Range demarcating line. The Allelic frequency was found out by gene counting. The population's pair-wise distance was calculated using Arlequin. 17 Y-STRs were collected from the data and the data was further compared to other similar studies using the minimal 9 Y-STRs standard by YHRD.

This study showed that the Northern population was genetically heterogeneous while the south population was genetically homogenous. Out of 118 haplotypes found in the 189 unrelated males, only 8 were shared between the two populations. The 9 Y-STR standard search resulted in similar haplotypes from Western Europe, India, Finland and Americas.

The study [3] was performed on the Pashtuns of Afghanistan. The Y-STR analysis of the Pashtuns of Southern Afghanistan was carried out in this study. Samples from 125 unrelated males of Southern Afghanistan were typed for Y-STRs which were then analyzed using YHRD online service.

As a result of this study it was observed that there is a close resemblance between the Pashtuns from Peshawar, (Pakistan) and Northern Afghanistan. Further-more the Pashtuns of South Afghanistan and India were quite distant from Northern Afghanistan and Pakistan.

The study [4] was carried out in 2002 to find out the genetic variation in Pakistan. DNA from 718 unrelated males was typed for 16 multi-allelic Y-STRs. 11 Haplogroups and 503 haplotypes were identified. The different Oral traditions and claims

were also scrutinized and this study nullified the claims of Pashtuns but also stated the considerable margin of error.

The limitation of these studies is that these studies were mainly conducted for a sub-population and not for the whole Pashtun population. Furthermore they were not carried out in consideration to each other and were more of individual studies. The origin of Pashtuns was also not addressed and when done then only for a single claim in one of the study. The oral traditions and claims were not addressed collectively in a single study. The different studied were also not collaborated to cumulate the findings and to reach to a single collective conclusion. Some of the studies were carried out almost 10 years back and there had been a lot of increase in YHRD database since then so further digging into the matter can bring up new dimensions and findings.

3 Data Analysis

The data collected was used in this study and consisted of 711 unrelated males from 12 ethnically different (populations) of Pakistan. The data from the 711 unrelated males was then typed for 16 Y-STRs. A total number of 527 haplotypes were found for all 711 individuals. In this study, the data for the Pashtun population consisting of 94 individuals was separated and analyzed on its own. The 73 haplotypes found among the 94 Pashtun males were typed. The 16 Y-STRs typed for this study were DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS425, DYS426, DYS434, DYS435, DYS436, DYS437, DYS438 and DYS439.

The Analysis of Molecular Variance (AMOVA) was performed on the data using the Analyze tool of YHRD. The F_{st} (fixation index) values for the data were calculated which are used to measure the genetic distance between different populations. The measuring of the genetic distance tells us whether the population samples are genetically similar and whether they could be joined together as similar populations or not. AMOVA for the Y-Chromosomal profiling is done. This method considers the number of times the Y-STR markers are repeated at the DYS (DNA Y-chromosomal Segment) loci. This is done both for inter population and intra populations. Different tests were carried out for the different pair-wise comparisons. The data from the Pashtun population was compared to the studies performed earlier whose data has been submitted to the YHRD database such as the Pashtuns of Northern and Southern Afghanistan, Afridi Pathans of India, Israel and Palestine, Macedonia, Western and Eastern Europe, Greece and limited African studies. Many population studies performed earlier were taken into account so as to narrow down the MRCA and verify the validity of the various oral traditions and claims.

The study as earlier stated was carried out to dig more into the origin of the Pashtun population. To cumulate the findings of the earlier studies (from the YHRD database) and to cumulate the research in a single study so as to conclude collectively from the earlier studies about the origin of Pashtuns or at least to take the research about the origin further on a notch.

The inclusion of all the 16 Y-STRs as a single set didn't lead to any conclusion because none of the data overlapped for all the 16 Y-STRs. The minimal YHRD

standard of 9 Y-STRs were then considered. The most common haplotype among the Pashtun population was searched for matches and no exact match was found with the minimal 9 Y-STRs and no nearest neighboring single step-mutation haplotypes were found. Similar searches were then performed for other sets of randomly selected haplotypes of the Pashtun population, but no matches were found. The STR marker count was then decreased to 7 Y-STRs. The 7 Y-STR markers were used to search for matches and used for analysis and Multi-Dimensional Scaling (MDS) plots were plotted using the YHRD. The MDS plot is a visual approach to the proximities pattern that the objects being tested have. The least the value means the closer the objects correspond to each other. This plot basically shows us graphically the closeness and distance between the objects under observation.

Haplogroup Prediction was also carried out for different haplotypes of the Pashtun population. The Haplogroup Prediction provided a further insight to the possible MRCA without having exact matches and keeping in view the single step-mutation.

4 Results

The Pashtun sample contained of 94 unrelated males with 73 haplotypes. The most common of these haplotypes was shared by 10 individuals with one shared by 6 and 2 sharing another haplotype. The search for the most common haplotype of the Pashtun population sample returned no exact match from the database. The markers were then reduced to 7 for the same haplotype and 13 exact matches were found from the Eurasian Meta-population (Eastern Europe) and 2 matches from Semitic population. Another common haplotypes search result showed 11 exact matches from Eurasian population and 2 from Indian population. 2 Iranian population exact matches were also found out.

Table 1. Fst Values for Scenario 1

Population	Afghanistan [Pathan]	Greece [Greek]	Israel & Palestinian	Brazil [African]	Afghanistan	Greece	Israel	Macedonia	Russian Federation	test2
Afghanistan [Pathan]	-	0	0	0	0.0002	0	0	0	0	0
Greece [Greek]	0.0211	-	0	0.0001	0	1	0	0.0001	0.0005	0.0027
Israel & Palestinian [Arab]	0.0229	0.0029	-	0.0002	0	0	1	0	0	0.002
Rio de Janeiro, Brazil [African]	0.0219	0.0018	0.0033	-	0	0	0.0001	0	0.0002	0.002
Afghanistan	0.008	0.0177	0.0197	0.0186	-	0	0	0	0	0
Greece	0.0218	-0.0026	0.0038	0.0026	0.0185	-	0	0	0	0.0024
Israel	0.0229	0.0029	-0.0062	0.0033	0.0197	0.0038	-	0	0	0.0009
Macedonia	0.0238	0.0024	0.0053	0.0041	0.0205	0.0025	0.0053	-	0	0.0018
Russian Federation	0.0217	0.0021	0.0051	0.0037	0.0168	0.0032	0.0051	0.0056	-	0.0015
test2	0.0222	0.0019	0.0035	0.0021	0.0188	0.0028	0.0035	0.0044	0.004	-

These results were in accordance with the earlier studies performed, confirming the most established opinion of the Eurasian decent of the Pashtuns. The Semitic trace further backs up the point of the Greek influence.

The Haplogroup prediction was carried out for the Pashtun population. The haplotype prediction resulted in R1a, J2, and L to be the possible Haplogroups. The Haplogroup percentages for the different Haplogroups were 61.6 % to be for the R1a, 31.6

% for the L and 6.7 % for J2 Haplogroup. The Haplogroup search resulted with L to be of least magnitude. The Haplogroup J2 was second highest in percentage. The J2 shares the African common ancestor and is linked to the Semitic population. This is found 46.1 % in the Greek population and 18.6 % in the Israel population.

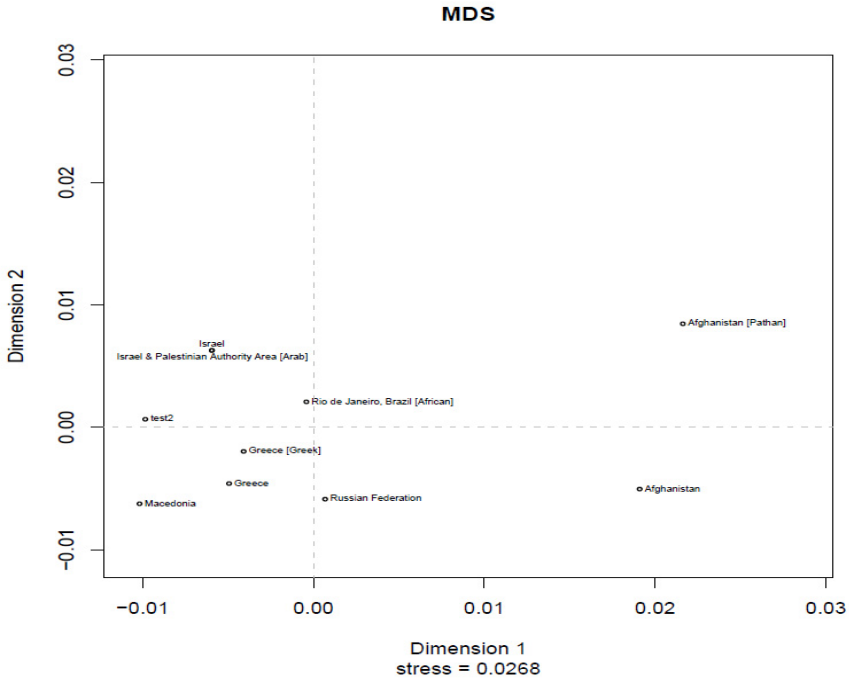


Fig. 1. MDS plot of scenario 1

AMOVA calculations and MDS plots were carried out for researches from the past and the current study collectively to cumulate the results in a single study. The Pashtun sample was then used for proximity search with any past population and data of the past researches.

A scenario for the correlation was that of the original data of Pashtuns with all the haplotypes used for the AMOVA calculation. The populations used in this scenario were Greece, Macedonia, Israel, Russia and Brazil. Table 1, shows the Fst values for the populations. The MDS plot for the same scenario shows the proximities of the populations.

Scenario two had the data formatted. Only two haplotypes from the Haplogroup 21 were found and neglecting these analyzed 5 populations and 5 population sets, which is a total of 7744 haplotypes which resulted in a further close relation to the Greece, Russian, Israel and Macedonian populations with all these populations along with our test data considered as a single cluster with the Pashtuns population of Pakistan as close neighbor. The Fst values were calculated which were used for the AMOVA calculation for the populations in the current scenario. The Fst values are shown in the Table 2 are the least values showing the highest correlation.

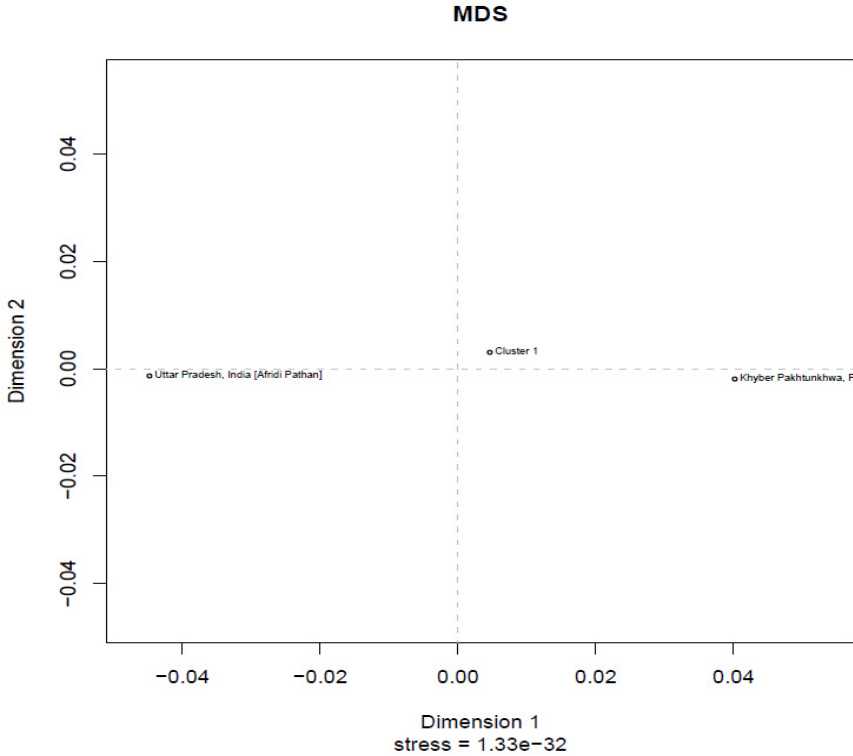


Fig. 2. MDS plot of scenario 1

Table 2. Fst values for different populations of scenario 2

Population	Afghanistan [Pathan]	[Yousafzai Pathan]	Northern Gre	South Afg	Uttar Prade:	Greece	Iran	Israel	Macedonia	Pakistan	TEST
Afghanistan [Pathan]	-		0	0	0	0	0	0	0	0	0
Pakistan [Yousafzai Pathan]	0.0555	-		0	0	0	0	0	0	0	0
Northern Greece, Greece [Greek]	0.0232	0.0426	-		0	1	0	0	0.0002		0.0014
South Afghanistan [Pathan]	0.035	0.0649	0.036	-		0	0	0	0	0	0
Uttar Pradesh, India [Afridi Pathan]	0.0555	0.0849	0.0552	0.0715	-		0	0	0	0	0
Greece	0.0218	0.0413	-0.0026	0.0342	0.0534	-		0	0	0	0.0028
Iran	0.0216	0.0409	0.0035	0.0335	0.0531	0.0024	-		0	0	0.0024
Israel	0.0229	0.0427	0.0049	0.0359	0.0548	0.0038	0.003	-		0	0.0013
Macedonia	0.0238	0.0434	0.003	0.0364	0.0553	0.0025	0.0046	0.0053	-		0.0015
Pakistan	0.0207	0.0247	0.0055	0.0322	0.0526	0.0045	0.0038	0.0058	0.0065	-	0.0011
TEST	0.0218	0.0422	0.0036	0.035	0.0546	0.0025	0.0022	0.0031	0.0041	0.0043	-

The MDS plot for the same scenario, given in Figure 1, shows the closely proximate populations as a single cluster with normalized stress value.

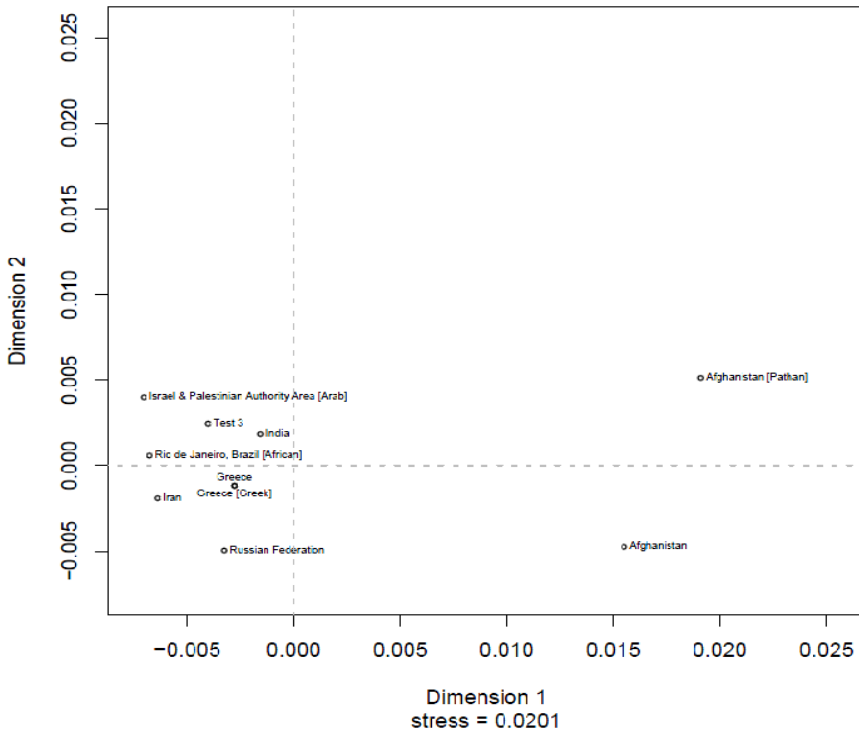


Fig. 3. MDS plot

The AMOVA calculations and the MDS plots show that the Pashtun population is closely proximal to the Greece, Russian Federation, Israel, Iran some Sub-Saharan populations. Most of the exact matches were from the Eurasian population (mainly Greece and Macedonia) with many exact matches from the Russian federation. The 7 Y-STR marker AMOVA resulted in exact matches from India and Iran. Exact matches from the Israel population only resulted for the 5 Y-STR marker test. The AMOVA calculations and the MDS plots show that the Pashtun population is closely proximal to the Greece, Russian Federation, Israel, Iran some Sub-Saharan populations. Most of the exact matches were from the Eurasian population (mainly Greece and Macedonia) with many exact matches from the Russian federation. The 7 Y-STR marker AMOVA resulted in exact matches from India and Iran. Exact matches from the Israel population only resulted for the 5 Y-STR marker test.

AMOVA calculation and the MDS plots showed that the Pashtuns of Pakistan had a close relation to the Pashtun Populations of Afghanistan and Populations of Greece, Europe, Macedonia, Russia and some African populations.

5 Conclusion and Discussion

The AMOVA calculations, MDS plots and the exact match results showed that the Pashtun population for 16 Y-STR markers did not provide any exact match but was in

proximity to the claimed ancestry of Greece and Aryan population. The exact matches from the Greece and Russian populations also back up the possible aforementioned decent. However most of the exact matches being from the Greece and Macedonia (Eurasian Meta-population) suggest strongly the decent from the Greece population.

The Haplogroup prediction of R1a, J2 and L showed that the results were in accordance with the previous studies. The R1a is the Eurasian population Haplogroup as anticipated. This as the Haplotype search also backed up as many exact matches from the same Haplogroup were resulted. Same Haplogroup is found in 8.3 % to 16.5 % in different studies in the Greek population. This backs the previous claim of Greek ancestry.

The Haplogroup J2 is from the Semitic population as well as the population sharing the common African ancestor. This Haplogroup was found 6.5 % in both the Greek and Pashtun population while 55.6 % in the Israel population. The Israel population however did not result in exact match for haplotype of the 9 or 7 markers tested. Very few exact matches were found only with the 5 markers test. However the 7 marker test had many exact matches from the Greek population.

On the basis of the results in hand, it can safely be said that the Israeli oral tradition can be regarded as a speculation as the results here show otherwise. With no exact matches and very different Haplogroup percentages, the Israel population descent can be stated as a myth.

The Greek population on the other hand showed similar Haplogroup percentages and many exact matches were found, so this study further backs the point of the Greek lineage of Pashtuns.

References

1. Fernandes, A.T., Gonçalves, R., Gomes, S., Filon, D., Nebel, A., Faerman, M., Brehm, A.: Y-chromosomal STRs in Two Populations From Israel and The Palestinian Authority Area: Christian and Muslim Arabs. *Forensic Science International: Genetics* 5(5), 561–562 (2011)
2. Ali, N., Coulson, T.: Announcement of Population Data: Genetic data for 17 Y-STR AmpF ℓ STR \circledR Yfiler TM markers from an immigrant Pakistani population in the UK (British Pakistanis). *Forensic Science International: Genetics* 7(2), 40–42 (2013)
3. Roewer, L., Willuweit, S., Kruger, C., Nagy, M., Rychkov, S., Morozowa, I., Naumova, O., Schneider, Y., Zhukova, O., Stoneking, M., Nasidze, I.: Analysis of Y Chromosome STR Haplotypes in The European Part of Russia Reveals High Diversities But Non-Significant Genetic Distances Between Populations. *International Journal of Legal Medicine* 122(3), 219–223 (2008)
4. Sammon, J.W.: A Non-Linear Mapping for Data Structure Analysis. *IEEE Transaction on Computers* 100(5), 401–409 (1969)
5. Achakzai, N.M., Rahman, Z., Shahzad, M.S., Daud, S., Zar, M.S., Israr, M., Husnain, T., Willuweit, S., Roewer, L.: Y-chromosomal STR Analysis in The Pashtun Population of Southern Afghanistan. *Forensic Science International: Genetics* 6(4), 103–105 (2012)
6. Lacau, H., Bukhari, A., Gayden, T., Salvia, L., Regueiro, M., Stojkovic, O., Herrera, R.J.: Y-STR Profiling in Two Afghanistan Populations. *Legal Medicine* 13(2), 103–108 (2011)

7. Mohyuddin, A., Ayub, Q., Qamar, R., Zerjal, T., Helgason, A., Mehdi, S.Q., Tyler, S.C.: Y-chromosomal STR haplotypes in Pakistani Populations. *Forensic Science International* 118(2-3), 141–146 (2001)
8. Robino, C., Varacalli, S., Gino, S., Chatzikyriakidou, A., Kouvatsi, A., Triantaphyllidis, C., Di Gaetano, C., Crobu, F., Matullo, G., Piazza, A., Torre, C.: Y-chromosomal STR Haplotypes in A Population Sample from Continental Greece, and the Islands of Crete and Chios. *Forensic Science International* 145(1), 61–64 (2004)
9. Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., Mehdi: Y-chromosomal DNA Variation in Pakistan. *The American Journal of Human Genetics* 70(5), 1107–1124 (2002)
10. Mansoor, A., Mazhar, K., Khaliq, S., Hameed, A., Rehman, S., Siddiqi, S., Papaioannou, M., Cavalli-Sforza, L.L., Mehdi, S.Q., Ayub, Q.: Investigation of The Greek Ancestry of Populations from Northern Pakistan. *Hum Genetics* 114(5), 484–490 (2004)
11. Firasat, S., Khaliq, S., Mohyuddin, A., Papaioannou, M., Tyler-Smith, C.: Y-chromosomal Evidence for A Limited Greek contribution to the Pathan population of Pakistan. *European Journal of Hum Genetics* 15(1), 121–126 (2007)
12. Aliev, A.A., Smirnoy, A.S.: Y Haplogroups of Carriers of Aryan Language. *The Russian Journal of Human Genetics* 2(1) (2010)
13. Excoffier, L., Lischer, H.E.L.: Arlequin suite ver 3.5: A New Series of Programs to Perform Population Genetics Analyses Under Linux And Windows. *Molecular Ecology Resources* 10, 564–567 (2010)

Two Notes from Experimental Study on Image Steganalysis

Qingxiao Guan^{1,2}, Jing Dong², and Tieniu Tan²

¹ Department of Automation, University of Science and Technology of China

² National Laboratory of Pattern Recognition, Institute of Automation
P.O. Box 2728, Beijing, P.R. China, 100190

Abstract. In recent years, several advanced methods for image steganalysis were proposed. During research process, some concerns are more and more addressed by steganalyzer. In this paper, we focus on several of these concerns. The first one is how to utilize SVM classifier in practical steganalysis, we use clustering analysis to divide training samples and train several SVM for detecting stego image. In this part we also discussed building an image database that can be used for evaluating steganography/steganalysis fairly. The second is how to designed proper classifier for steganalysis, especially how to take information of cover/stego image pair into account. We will discuss several notions regard to these two concerns.

Keywords: Steganalysis, steganography data base, clustering, feature analysis.

1 Introduction

As the media files spread rapidly on internet, some vicious and illegal message may reside in innocent-looking images. Secret message can be hidden in images by very small modifications on pixel gray value or JPEG coefficients. Such images, which were modified by steganography methods, were called stego image. Stego image will not arouse any suspicious by human eyes. Thus it is better to detect stego images by computer algorithms. Steganalysis methods were developed to detect such images and became a key issue in digital forensic field. From previous literature, steganalysis can be categorized into two kinds: specific and universal. The former can deal with certain kind of steganography and utilize special prior-knowledge of steganographic scheme in detection. The later, on the other hand, do not require any prior or special information of a particular steganography, but only by analyzing the inherent characteristics of cover media data, thus is more universal for practical user. Recent developed steganalysis methods are more likely to be the later kind. In this paper we mainly focus on two important concerns for universal image steganalysis. Without confusion, in the rest of this paper, the word "steganalysis" is used specifically for universal image steganalysis.

Steganalysis typically includes two parts: extracting features and using classifier to predict result from feature. Features usually are some statistics calculated from image. It is basic information for steganalysis. And the classifier is trained and used for mapping features to the prediction result. There are several proposed feature sets[1, 2, 3],

and modern classifiers (such as SVM, or assemble classifier) were introduced to this field. The topic of this paper involved in both aspect, and meanwhile related to the characteristic of steganalysis.

The first problem is about using SVM in practical steganalysis. SVM with RBF kernel was used detecting many kinds of stego images in steganalysis. However it suffered unacceptable computational and storage cost when training on large sample set and testing a candidate image, which is very important to obtain generalize ability for dealing with large variance of image content in practical applications. For the reason that steganography embed secret message in image only by slightly modification, the quantitative properties of steganalysis feature is less disturbed by embedding, but greatly influenced by image content. This problem brought difficulties for steganalysis due to the complexity of nature image content. In this paper we proposed a practical method to solve this problem. When detect a candidate image by inputting it into classifier, we assumed that detection ability of RBF SVM was provided by training samples which are close to candidate image in feature level. Thus we propose a novel steganalysis scheme based on clustering and RBF SVM in this paper. It first divides training sample set into subsets by clustering, then training classifiers on samples in each subset. In testing procedure, we measure the distance between feature of candidate image and centroid of subsets. Then select a classifier by distance measures to detect candidate image. In this part we also investigate construction of database used for evaluation on steganalysis performance. In previous literatures, the most objective way to evaluating a steganalysis method is carrying out experiment on a standard database. However, if we want to compare two steganalysis algorithms, many factors may affect the performance. For this topic, most people focus on processing of image, such as rescale or compress. We can see the performance of SPAM[1] on different database varies so much. Similar conclusion can be drawn from other experiments. However, in most literatures, the experiment was only implemented on single database, or mixed database. Even for certain database of identically processed images, one of the biggest problems is the complexity of image content. In another word, accuracy of detection in steganalysis varies greatly to different content. Unfortunately, except in [4], we found no more work discussed this problem. Compare to affection of processing, this problem is much more difficult to study. Although enriching variety of image in database is a common solution to avoid bias, it is almost impossible to declare an evaluation is absolutely fair, because we can only perceive image content, not its influence to micro level of statistics used by steganalysis. Most databases used in steganalysis include images of several categories such as indoor, building, animal and plant. In steganalysis, even in the same category, features of two original images (without data hiding) may differ very much. Instead of discussions with ambiguous concept, we try to reveal some properties of this aspect by statistical analysis. Thus in section 2 we show some experimental results on BOSS database, which is direct evidence that support our viewpoint. Moreover, this problem pushes us to consider another criterion of evaluating the "fairness" of dataset for steganalysis. Alternatively, we analyze this problem from distribution of feature. The first part of this paper is our discussions on this topic. We use K-means clustering to heuristically justify some basic assumptions, and based on that we proposed a model for ideal database.

The second aspect is how to design proper classifier for steganalysis. It is still an open problem and we just make some remarks and give some possible ideas regarding

this topic in this paper. Previous classifiers used in steganalysis are always "problem-independent". They are not specialized for steganalysis. This may partly due to the difficulties of explicitly describing problems in steganalysis. Since special knowledge in certain applications may help to enhance classifier, we just try to make use of some knowledge for training the classifier. This idea is also originated from some observations in recent experiments. There is a requirement for this training scheme: training samples include Cover/Stego pairs (Abbreviate as C/S-pair). C/S-pair is an image pair composed by a cover image and a corresponding stego image. In each C/S-pair, the stego image is generated by cover image in the same C/S-pair. The relationship of images in a C/S-pair is precious information and should be used for training the classifier. Compare to other pattern recognition problem, the formulation of such C/S-pairs is a uniqueness of steganalysis. This is the reason we seek to design a learning mechanism which involves this information. In fact, the requirement of providing C/S-pair in training process can be easily satisfied, because they can be prepared by steganalyzer. We must mention that similar ideas of C/S-pairs also appear in [5], but they only use C/S-pairs in cross validation of ordinary SVM.

Although we are still studying the above two important concerns for steganalysis, we are willing to present our ideas and discuss some possible solutions or suggestions for better dealing with their problems.

The rest of this paper is organized as follows: In Section 2, we discuss how to build a database to evaluate each steganalysis system fairly, we also present some inferred reasons from our experiment result. In section 3, we present how to use C/S-pairs in Fisher Linear Discriminate(FLD) and SVM classifiers. Section 4 is the conclusion part.

2 Practical Steganalysis by Clustering and RBF SVM

RBF SVM is the most popular classifier in steganalysis for its high accuracy and robustness[8]. Since RBF SVM model for detecting stego image is obtained by training on training set, it is important to adding more samples to training a model with high generalize ability. However, more training samples bring great burden in training process due to the large number of support vectors. To solve this problem and meanwhile keep high accuracy in steganalysis, we first divide training sample set by clustering. After clustering, samples with similar feature were aggregated in the same subset. Thus we can train classifiers individually to avoid training on the whole training set. In testing procedure, feature is extracted from candidate image and compared with centroids of subsets $\|x - \mu_k\|_2, k=1,2,3...c$. The distance between centroid of subset and feature of candidate image present their "closeness". Classifier of the closest subset is assigned to detect candidate image:

$$\begin{aligned} result &= classifier_{k^*}(x) \\ k^* &= \arg \min_k \|x - \mu_k\|_2 \end{aligned} \quad (1)$$

Where k denotes the k th subset of training samples after clustering, μ_k denotes the centroid of k th subset, x denotes feature of candidate image, and $classifier_{k^*}(\bullet)$ is classifier trained by k th subset of training samples.

We use K-means clustering method for dividing training set. K-means clustering is a classic clustering method and can be used for large scale data for its simplicity.

Database always play a very important role in evaluation. There are several database used by steganalyzer to evaluate their steganalysis algorithms. Among these databases, BOSS[6] is the most standard, and has largest number of images captured by 6 cameras. We start our discussion in this section from a very simple experiment on BOSS database.

The experimental setting is described as follows: BOSS database v1.0 includes 10000 images, and we embed message into these images by Hugo steganography algorithm[7] with 0.1 and 0.2 bpp (bit per pixel) payload to create 10000 0.1bpp stego images and 0.2 bpp stego images. We extract NIP feature[3] from these cover and stego images and use K-means cluster algorithm to cluster these samples into 5 clusters. NIP feature is a feature set for steganalysis. In [3] the author only use first order subtraction to formulate NIP feature, but in this paper we extend NIP feature to more types using residuals proposed by Kodovsky et al in [2]. These residuals include first and second order subtraction residual, KB residual, EDGE residual and MAX-MIN residual. They can be applied to NIP feature by replacing the first order residual in NIP feature with them. More types NIP feature enlarged the scope of representing stego image and consequently enhanced the ability for detection. Our feature in this paper is formed by combining NIP feature of these types.

$$\begin{aligned}
1 \text{ st} \quad D_{i,j}^{1,\uparrow} &= X_{i+1,j} - X_{i,j} \\
2 \text{ nd} \quad D_{i,j}^{2,\uparrow} &= X_{i-1,j} - 2X_{i,j} - X_{i+1,j} \\
3 \text{ rd} \quad D_{i,j}^{3,\uparrow} &= X_{i-1,j} - 3X_{i,j} + 3X_{i+1,j} - X_{i+2,j} \\
EDGE \quad D_{i,j}^{EDGE,\uparrow} &= -X_{i-1,j-1} + 2X_{i,j-1} - X_{i+1,j-1} \\
&\quad + 2X_{i-1,j} - 4X_{i,j} + 2X_{i+1,j} \\
KB \quad D_{i,j}^{KB} &= -X_{i-1,j-1} + 2X_{i,j-1} - X_{i+1,j-1} \\
&\quad + 2X_{i-1,j} - 4X_{i,j} + 2X_{i+1,j} \\
&\quad - X_{i-1,j+1} + 2X_{i,j+1} - X_{i+1,j+1}
\end{aligned} \tag{2}$$

For each pixel, we compute residual of 4 directions, for example using 1 st residual to compute neighbor information on pixel (i, j) :

$$d_{i,j} = \{D_{i,j}^{1,\uparrow}, D_{i,j}^{1,\rightarrow}, D_{i,j}^{1,\downarrow}, D_{i,j}^{1,\leftarrow}\} \tag{3}$$

For two pixels in position $(i1, j1)$ and $(i2, j2)$, if $d_{i1,j1}$ and $d_{i2,j2}$ are the same, or can be convert to the same by rotating one of them by 90,180,or 270 degree, we say they are “rotation invariant” neighbor information. Combining these “rotation invariant” neighbor information and calculating its histogram, neighbor information feature is obtained.

Please note that the clustering is an unsupervised learning method, which means the label of each sample is not used and every sample is treated equally. K-means clustering assign each sample into a cluster just by the distance between feature and cluster center, and then revise centers of clusters. These two steps are iteratively

repeated until converge. After clustering, the samples are divided into clusters. In each cluster, we randomly select 80% samples to training a classifier, and rest 20% samples for testing. In this experiment, testing samples are classified by the classifier corresponding to the cluster. We also compare this scheme with conventional scheme which use all samples without clustering to train a single classifier. In fact, we do not specifically select parameters for each cluster in training classifier, but only use parameters for training all the classifiers. The Table 1 and Table 2 is the result of this experiment on Hugo 0.1bpp and Hugo 0.2bpp. The first column is cluster tag and columns 2 to 4 are False Alarm rate(FA), Miss Detection rate(MD), and Error rate(Error) respectively. FA denotes the portion of incorrectly classified cover samples in testing cover samples. So MD denotes the portion of incorrectly classified stego samples in testing stego samples. Error is the overall error rate for all testing sample. Column 3 to 5 is time cost for training and testing. Column 6 to 7 is number of training samples and testing samples for each cluster. And the last 2 rows are overall performance of our scheme and conventional scheme. "All cluster" is overall rate of our scheme which uses clustering. And "All sample" is conventional scheme of training single classifier without clustering.

Table 1. performance of detecting stego image embedded with 0.1bpp by Hugo

Clusters index	FA (%)	MD (%)	Error (%)	Training time (s)	Testing time (s)	Training sample number	Testing sample number
1	42.95	38.00	40.48	141.8	46.7	3628	908
2	45.80	49.20	47.50	157.8	68.3	4272	1070
3	17.24	14.22	15.73	61.2	7.0	1848	464
4	34.62	30.95	32.79	303.4	52.3	4132	1034
5	50.67	45.28	47.98	33.3	17.1	2114	530
All clusters	39.66	37.43	38.54	697.5	191.4	15994	4006
All samples	39.85	36.85	38.35	113905.3	710.1	15994	4006

Table 2. performance of detecting stego image embedded with 0.2bpp by Hugo

Cluster index	FA (%)	MD (%)	Error (%)	Training time (s)	Testing time (s)	Training sample number	Testing sample number
1	31.06	26.21	28.64	181.2	34.2	3630	908
2	56.00	43.94	49.97	36.5	17.0	4278	1070
3	3.02	6.04	4.53	14.9	2.4	1844	464
4	44.49	38.32	41.41	230.1	63.9	4132	1034
5	16.25	16.44	16.35	228.7	28.1	2110	530
All clusters	30.89	26.93	28.91	691.4	145.6	15994	4006
All samples	29.9	25.1	27.5	56225.5	692.7	15994	4006

From Table 1 and Table 2, we can see that, although we did not use any class information of samples in clustering, accuracy varies so much between different clusters. Some clusters that include less samples have very high accuracy, and seems do not suffered from “dimension of curse”. On another hand, the total accuracy is very similar to the scheme of using all the training samples for training (without clustering). But the time cost of our method is much lower, which implies our method is promising for construct practical steganalysis system. We also observed that, after clustering each cover image and its stego images stayed in the same subset. This is not surprising because they have very small difference caused by steganographic embedding and main quantitative character is determined by image content. From this experiment we make an assumption: in detection, discriminative information for a testing sample is mainly provided by those training samples in the same cluster. And the anti-detection ability of samples in different clusters is different. We checked images in each cluster, and find that clusters contain images of different content captured by different cameras. From this point, deliberately diversify image content in database can hardly avoid imbalance of each cluster and may cause a bias result in testing.

Supposing the whole image database contains N images, and N is large enough. They were cluster into K clusters, and the i th cluster contains N_i images, and for each cluster the detection error rate is $Error_i$. Consequently the total error rate $Error_{cluster}$ is:

$$Error_{cluster} = \frac{\sum_{i=1}^K Error_i \times N_i}{\sum_{i=1}^K N_i} \quad (4)$$

We can see that $\frac{N_i}{\sum_{i=1}^K N_i}$ plays as a weight factor in $Error_{cluster}$.

Since $Error_i$ vary between different i , the number of each cluster affects $Error_{cluster}$. We also use $Error_{all}$ to denote the accuracy of detection scheme which use all the training samples in training phase. As we assumed, if N is large enough, $Error_{cluster}$ is an approximation of $Error_{all}$. Under this assumption, $Error_{all}$ in performance evaluation partially determined by the number of test samples assigned to each subset.

Based on experimental result and analysis, now we present the model for an ideal database. If we have enough image samples, after clustering by any kind of feature with any setting of cluster number K , each cluster includes the same number of samples. If the database is consistent with this principle, we believe that it is an ideal database without any bias for testing features. Selecting samples for building a database is just like a “sampling” process, and this principle would guarantee that the database is a uniform “sampling” for any kind of feature.

We propose this principle based on our experiment, besides above described experiment, we also did some experiments on SPAM feature[1] and received similar results that support our viewpoint. We did not test that using other higher dimension feature such as Holmes[2], because dimensionality of Holmes is very high and the number of images in BOSS database is limited. We just used the K-means clustering method to demonstrate our basic idea. And for other clustering methods, we believe the result will be similar if the number of samples and clusters is large enough. This is an arguable conclusion.

In practice, many factors may affect the final result, but we only consider a single factor. This model is unable to give any guidance of enlarging image diversity in database. But for another case, if we have sufficient images as candidates, we can select a subset from them by this model, and then evaluating steganalysis performance by this subset to get a relatively fair result.

3 Using C/S Pairs for Training Classifier

Classifier is an indispensable part in steganalysis system. It maps the feature to its prediction result. The classifier is generally trained by training samples. However, classifiers always consider the most general conditions. In this section we propose some ideas of design classifiers that take C/S pairs into account.

As we mentioned in previous sections, for a test sample, the most discriminative information was provided by those samples nearest to it in terms of feature distribution. For the convenience, we first define some notations: let S_{train} and S_{test} denote the set of training samples and testing samples. The conception of "sample" is equal to feature of sample. For a sample f , the $N_k(f)$ denote the set of nearest k samples of f in S_{train} . In the definition of $N_k(f)$, the distance of two samples can be l_1 or l_2 norm. And a C/S pair is denote as a tuple (c, s) , c is a cover sample, after embedding message in c , s is obtained. Based on above notation system, we further define a special set of C/S pairs:

$$A_k = \{(c, s) \mid \exists f \in S_{test}, c \in N_k(f) \text{ or } s \in N_k(f)\} \quad (5)$$

A_k includes C/S pairs in raining set. Each C/S pair in A_k at least contains a sample that belong to $N_k(f)$ of a sample $f \in S_{test}$.

We first present an experiment. This experiment is a testing of detection. We use BOSS database[6], and Hugo 0.4bpp stego images. BOSS include 9047 images, thus we have 9047 C/S pairs. We randomly select 20% C/S pairs as testing samples, and the rest used for training. In this experiment, we use Holmes feature sets and random subspace method[2] proposed by Kodovský et al. Holmes feature set and random subspace are so far the most effective feature set and classifier. For each C/S pairs, the testing result can be ordered in 3 cases:

Case 1: Two images in the C/S pair are correctly classified.

Case 2: Only one image in the pair is correctly classified, in another word, two images in the C/S pair are classified in the same class.

Case 3: Two images in the pair are incorrectly classified.

Finally the result is interesting: 33.2% C/S pairs in testing set belong to Case 2. And on contrast, C/S pairs Case 3 only account for 0.8%. We repeat this experiment several times and the results are similar. Inspired by this result, we formulate some criterions for training classifier with C/S pair. We discuss two prototype of classifier: FLD and SVM, and show how to use C/S pairs in training FLD and SVM.

The first one is FLD. Briefly review the criterion of training FLD:

$$\alpha^* = \arg \min \frac{\alpha^T S_W \alpha}{\alpha^T S_B \alpha} \tag{6}$$

Where S_W and S_B is inner-class and intra-class matrix respectively. And α^* is weights of linear classifier. The meaning of this function is obvious: minimize inner distance of each class under certain intra-class distance. In this function all the samples are considered equally. We add another factor to this function to encourage classifier using those C/S pairs which near testing samples. The FLD is modified as follow:

$$\alpha^* = \arg \min \frac{\alpha^T S_W \alpha - \lambda \sum_{(c,s) \in A_k} \|\alpha^T c - \alpha^T s\|^2}{\alpha^T S_B \alpha} \tag{7}$$

Where (c, s) is pair in set A_k , and λ is a parameter of this factor.

The motivation of factor $\sum_{(c,s) \in A_k} \|\alpha^T c - \alpha^T s\|^2$ is explicit, and it is in a quadratic form, thus we can find a matrix P to rewrite it as $\alpha^T P \alpha$, and the function can be write in the form of:

$$\alpha^* = \arg \min \frac{\alpha^T (S_W - \lambda P) \alpha}{\alpha^T S_B \alpha} \tag{8}$$

This is still a generalized eigenvalue problem.

SVM is also a popular classifier, it maximize margin of support vectors. The model is found by solving:

$$\begin{aligned} \min_{w,b,\zeta} & \langle w, w \rangle + C \sum_{i=1}^l \zeta_i^2 \\ \text{s.t.} & y_i (\langle w, x_i \rangle + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, \dots, l \end{aligned} \tag{9}$$

Where x_i is the i th sample, w is decision boundary and b is bias.

After simple analysis, we noticed the factor $C \sum_{i=1}^l \zeta_i^2$ in the object function correspond to the ‘‘soft margin’’. This factor makes a trade-off between selecting

support vector and error of training samples. And C is a weight that determine how much important we attach to the correctness of classifying training samples. We can emphasize those C/S pairs in training process. This can be simply implemented by adjusting their weight:

$$\begin{aligned} \min_{w,b,\zeta} \quad & \langle w, w \rangle + C \sum_{i=1}^l \zeta_i^2 + C' \sum_{(x_i, x_j) \in A_k} (\zeta_i^2 + \zeta_j^2) \\ \text{s.t.} \quad & y_i (\langle w, x_i \rangle + b) \geq 1 - \zeta_i, \quad i = 1, \dots, l \end{aligned} \quad (10)$$

Where C' is the parameter which dominates the affection of C/S pairs. C' is positive. This factor do not break original criterion of SVM. And it can be solved without using special program. This is still a quadratic program with linear inequality constrains. In this model the parameter C' and k should be empirically selected.

4 Conclusion

In this paper we have presented some thoughts and discussions summarized from our experiments. We proposed a practical method to utilize high accuracy classifier RBF SVM in steganalysis of large data. And experimental result demonstrates that this method is simple and promising. Database construction is another key point in researching. Any existing database can hardly cover all the complex conditions in real application. We all want to evaluate steganalysis method as fair as possible based on certain database. There are some issues concerned by steganalyzers. One of the most severe problem is mismatching[4]. Some testing samples are misclassified because of lacking training samples captured from camera of the same type. In practice, when coupled with other problems, it is hard to assert an evaluation result fair or not. Our proposition of database construction aimed to balance those factors in an intuitive way. It requires a database with very large number of images. For this requirement, so far the most eligible database is BOSS V1.00, and we are looking forward to new versions. From the clustering experiment, we also find the time cost of clustering and training classifiers is much lower than convention training. So is the time cost of detection. We need more experiments to investigate it. Classifier is another important aspect in steganalysis. For particular problem, incorporating pre-knowledge into it can improve performance. We incorporate some pairwise information into designing classifier. This is a preliminary idea. Probably more intrinsic principals of steganography will be revealed in future. They can be employed in classifier design if they can be properly formulated. Classification design relies on empirical result. In our experiments we only considered some relationship of C/S pairs and proposed some possible forms. However, they can be utilized in other forms. We addressed two important issues in this paper. The improvement for these two issues will ensure the practical use for steganalyzers. Some are very tough for steganalyzer, such as cover source mismatching. We believe further works will still focus on feature and classifier design, and learning method will also become more significant.

References

1. Pevny, T., Bas, P., Fridrich, J.: Steganalysis By Subtractive Pixel Adjacency Matrix. *IEEE Transactions on Information Forensics And Security* 5(2), 215–224 (2010)
2. Fridrich, J., Kodovský, J., Holub, V., Goljan, M.: Steganalysis of Content-Adaptive Steganography in Spatial Domain. In: Filler, T., Pevný, T., Craver, S., Ker, A. (eds.) *IH 2011*. LNCS, vol. 6958, pp. 102–117. Springer, Heidelberg (2011)
3. Guan, Q., Dong, J., Tan, T.: An Effective Image Steganalysis Method Based on Neighborhood Information of Pixels. In: *Proc. IEEE International Conference on Image Processing, ICIP 2011* (2011)
4. Fridrich, J., Kodovský, J., Holub, V., Goljan, M.: Breaking HUGO-The Process Discovery. In: Filler, T., Pevný, T., Craver, S., Ker, A. (eds.) *IH 2011*. LNCS, vol. 6958, pp. 85–101. Springer, Heidelberg (2011)
5. Schwamberger, V., Franz, M.O.: Simple Algorithmic Modifications for Improving Blind Steganalysis Performance. In: *Proceedings of the 12th ACM Workshop on Multimedia and Security, MM&Sec 2010*, pp. 225–230. ACM (2010)
6. Bas, P., Filler, T., Pevný, T.: Break Our Steganographic System— The Ins And Outs of Organizing Boss. In: *13th International Workshop on Information Hiding* (2011)
7. Pevný, T., Filler, T., Bas, P.: Using high-dimensional image models to perform highly undetectable steganography. In: Böhme, R., Fong, P.W.L., Safavi-Naini, R. (eds.) *IH 2010*. LNCS, vol. 6387, pp. 161–177. Springer, Heidelberg (2010)
8. Mohammadi, F.G., And Abadeh, M.S.: A Survey of Data Mining Techniques for Steganalysis. In: Sajedi, H. (ed.) *Recent Advances in Steganography*, November 7 (2012)

Robust and Efficient Iris Recognition Based on Sparse Error Correction Model

Wei Cao^{1,2}, Yun Song^{1,*}, Zunliang He¹, and Zhimin Zhou²

¹School of Computer and Communication Engineering,
Changsha University of Science & Technology

Caowe_cn@163.com, sonie@126.com, zlhe@csust.edu.cn

²College of Electronic Science and Engineering, National University of Defense Technology
kdzhouzhim@yahoo.com.cn

Abstract. Iris recognition has become one of the most promising approaches for biometric authentication. Due to the fact that corruption and occlusion in iris images caused by eyelash occlusion, eyelid overlapping, specular and cast reflection is large in magnitude but sparse in spatial, a sparse representation method based on sparse error correction model is introduced in the paper. To improve the robustness and efficiency of the recognition system, each iris sample is separated into a few sectors, and a Bayesian fusion-based cumulative SCI (CSCI) approach is applied to validate the recognition results. Experimental results on CASIA-IrisV3 demonstrate the proposed method achieves excellent recognition performance both in robustness and efficiency.

Keywords: Iris recognition, Sparse representation, Sparse error correction, Cumulative sparsity concentration index.

1 Introduction

Matrix factorization can be used to deal with some kinds of learning problems. Different applications determine the cost functions and constraints imposed. Therefore, sparse representation of face image signals being modeled by matrix factorization has become a hot research topic in recent years. In [1], Donoho and Elad introduced an optimal sparse representation approach in general dictionaries(nonorthogonal) by l^1 minimization. Among the useful results stemming from the paper, one is very important, i.e, on condition that nonzero entry number of the l^0 -norm solution is less than $\frac{M}{2} + 1$, the l^0 -norm solution is the unique l^1 -norm solution, where M is a bound of all off-diagonal entries of the Gram matrix ($G = D^T D$) of the dictionary D . If the basis vectors of the dictionary are approximately orthogonal and the parameter M is very small, a satisfactory result for sparse representation was achieved. It becomes important to get a very sparse basis matrix or an approximately orthogonal one. Instead of using the generic dictionaries (with standard bases, e.g., Wavelet), the test sample coming

* Corresponding author.

from an overcomplete dictionary whose base elements are the training samples themselves was presented to apply to face recognition [2]. In this paper, we introduce a robust and efficient sparse representation method based on Sparse Error Correction model for iris image recognition. Considering the fact that errors incurred by occlusion, secularities and cast shadows are large in magnitude but sparse in spatial, it is more reasonable that a test sample is better represented by a combination of the training samples with additive sparse and large-magnitude errors than that used in the standard SRC model [3]. By effectively choosing the base vector which can most compactly express the input iris image and reject other possible but less compact samples, the proposed method shows excellent and robust recognition performance.

2 Process of Iris Recognition via SEC Model

Sparse representation approach has proven to be an attractive tool for some application domains such as iris recognition, which results from the fact that high dimensional iris images have the property of degenerate structure. According to the manifold learning method [4], an iris observation sample lies on a low dimensional manifold of the high dimensional iris images space. That is, an iris test sample can be sparsely coded by the training samples on the manifold. By means of solving the l^1 minimization problem, we can get the sparse coefficient vectors which exhibit explicitly the relationships among the iris training samples. For iris images, differing from face images, the corruptions and occlusions usually occur at some fixed regions, e.g., areas near upper and lower eyelids. Some measures will improve the robustness and efficiency of the recognition system greatly, such as slicing the iris image into a few blocks, discarding some highly corrupted or occluded sectors and combining some sectors of good quality [3]. In practice, we add an image separating step into the SEC Algorithm, which divides every iris image sample into K non-overlapping sectors of equal size, thus a dictionary is got and a corresponding classifier is learned for each sector respectively. Therefore, there will be K sparse coefficient sectors for each test sample. We make decision in term of the confidences of the sparse coefficient vectors.

3 Sparse Error Correction Model

Based on the manifold learning method, Wright et al. [2] reported a sparse signal representation approach named the SRC algorithm in face recognition tasks. That is, given n training samples $d_{i,1}, d_{i,2}, \dots, d_{i,n} \in \mathbb{R}^m$ of the i th class object, the new test sample $y_i \in \mathbb{R}^m$ of this object will approximately lie in the linear span of the training samples:

$$y_i = d_{i,1}x_{i,1} + d_{i,2}x_{i,2} + \dots + d_{i,n}x_{i,n} \quad (1)$$

where $x_{i,1}, x_{i,2}, \dots, x_{i,n} \in \mathbb{R}$ are the scalar coefficients. The iris recognition problem can be taken as seeking a sparse representation of the test iris in terms of the training iris images, just as the method of SRC for face recognition. The iris samples from the i th

subject are concatenated to form a matrix $D_i = [d_{i,1} | d_{i,2} | \dots | d_{i,n}]$. Thus any test image y can be written as

$$y = Dx \tag{2}$$

where x denotes a coefficient vector, $D = [D_1 | D_2 | \dots | D_k]$ is a matrix formed by concatenating the entire training set from all k classes. In (2), x is sparse and only the entries associated with the i th class of y are nonzero, i.e. $x = [0, \dots, 0, x_{i,1}, x_{i,2}, \dots, x_{i,n}, 0, \dots, 0]^T$. The special sparse structure of the coefficient vector is highly recognizable. Ideally, it identifies the label of the sample directly. However, in practical scenarios, the test samples are often partially corrupted or occluded. In this case, the low-dimensional linear model formulated in (2) should be modified to an error correction version [2]:

$$y = Dx + e \tag{3}$$

where $e \in \mathbb{R}^m$ is a vector of errors, namely, an error correction term for the low-dimensional linear model.

For robust recognition task, sufficiently large size of training samples makes equation (3) underdetermined with no unique solution. Under ideal condition, the solution of x is sparse enough if the number of the object classes is sufficiently large. Furthermore, considering the most common iris corruptions which characterize to be large-magnitude, sparse and spatially localized, the vector e is sparse and its non-zero items are large-magnitude. To minimize the energy of the errors, we propose the sparse error correction (SEC) model which can be formulated as:

$$(\hat{x}, \hat{e}) = \arg \min \|x\|_0 + \|e\|_0 \quad s.t. \quad y = Dx + e \tag{4}$$

where $\|\cdot\|_0$ denotes the l^0 -norm which counts the number of nonzero in the vector. The optimization problem in (4) is NP-hard, and it has been revealed that under certain mild conditions, (4) can be converted to the following convex optimization problem [1]:

$$(\hat{x}, \hat{e}) = \arg \min \|x\|_1 + \|e\|_1 \quad s.t. \quad y = Dx + e \tag{5}$$

where $\|\cdot\|_1$ denotes the l^1 -norm, i.e. $\|x\|_1 = \sum_i |x_i|$. A sparse solution can be found in polynomial time via Basis Pursuit (BP) algorithm [6, 7]. So we can get the coefficient vector x_i of the i th class, and we can identify the class of the test image by the index whose coefficients minimize the reconstruction residual:

$$\hat{i} = \arg \min \|y - D_i x_i - e\|_2 \tag{6}$$

where $\|\cdot\|_2$ denotes the l^2 -norm, \hat{i} is the recognized object class.

4 Validation Measures

In practical recognition system, there exist some situations which will influence recognition performance greatly, e.g., a new test sample could be an iris subject unauthorized, a test iris image is corrupted or occluded seriously, an image that is not an iris at all. Therefore, it is crucial for an iris recognition system to detect and reject suspect recognition results and invalid test samples. In [2], the notion of sparsity concentration index (SCI) was introduced to validate the recognition results. We adopt the SCI as the confidence of the recognition result. SCI with the coefficient vector x_i is defined as:

$$SCI(x) = \frac{k \cdot \max_i \left\| \prod_i(x) \right\|_1}{\|x\|_1} \quad (7)$$

where k is the number of the classes, x_i is the coefficient vector of the i th class and $\prod_i(x)$ denotes a vector retaining the labels associated with the i th class and setting others to zero. The value of SCI is between 0 and 1. $SCI(x) = 1$ means the test image is identified as a single class and the recognition result is highly confident. $SCI(x) = 0$ suggests the values of the coefficient vector are spread evenly over all classes and the test sample is likely to be invalid. A threshold between 0 and 1 can be chosen as the criterion for validating. If SCI is larger than the threshold, the recognition result is accepted. Otherwise, the test sample is treated as invalid. Different regions of the iris image acquired in practical scenarios have different qualities. Dividing the iris into several sectors, recognizing them separately and combining the results according to the quality of different sectors will greatly improve the recognition rate. Considering the recognition result confidence, in [3], a Bayesian fusion-based cumulative SCI (CSCI) approach is proposed to combine the recognition results of different sectors. We use CSCI as our result validation strategy. CSCI of the i th class is defined as the sum of the SCI values of all the sectors labeled as the i th class:

$$CSCI(i) = \frac{\sum_{j=1}^M SCI(c_j) \cdot \delta(c_j = i)}{\sum_{j=1}^M SCI(c_j)} \quad (8)$$

where M is the number of validate sectors whose SCI values are larger than the given threshold. The sectors with low confidence are discarded in advance. c_j denotes the recognition result of the j th block, i.e. the index of the largest entry of the coefficient vector of the j th sector. $SCI(c_j)$ represents the SCI value of the j th sector computed by (7). $\delta(\cdot)$ is the Kronecker delta function. Thus, if the CSCI values of all the classes are smaller than the threshold, the sample is rejected. Otherwise, the recognition result is given by $c = \arg \max_{i \in C} (CSCI(i))$, here C is the set of all the classes.

5 Experimental Results with the CASIA-IrisV3

From CASIA-Iris-Lamp subset coming from CASIA-IrisV3 [5], we randomly select 120 object classes and 20 iris images under different illumination conditions for each class. The code of Masek and Kovesi [8] is applied to segment the iris and each iris sample stacks as a 4800 dimension vector. Spectral Projected Gradient (SPGL1) algorithm [7] is employed to solve the BP problem for recovering representing coefficients. In the experiments, we set the SCI and CSCI rejecting threshold to 0.7. The iris samples were processed respectively by SRC and SEC method with a varying number of blocks and the feature dimension. Fig. 1 shows the performance of SEC is superior to SRC. The recognition rate improves significantly with the number of sectors increased from one to eight, which indicates that dividing the iris to eight sectors can exclude the corrupted or occluded region most effectively (Fig. 1a). It can be seen the recognition rate increases monotonically with the feature dimension (Fig. 1b). We classes (eyes) from CASIA-Iris-Lamp and assign them into three groups with 120 classes in each group according to iris samples of different quality. An iris sample is divided into eight sectors. The SCI and CSCI threshold are chosen evenly in the range [0, 1], e.g., [0, 0.1, 0.2, ... , 1]. SRC and SEC methods are employed experimenting on the original iris samples by changing the training number and feature dimension respectively. Table 1 shows that under same conditions the performance of SEC method gain over SRC for all images of different quality.

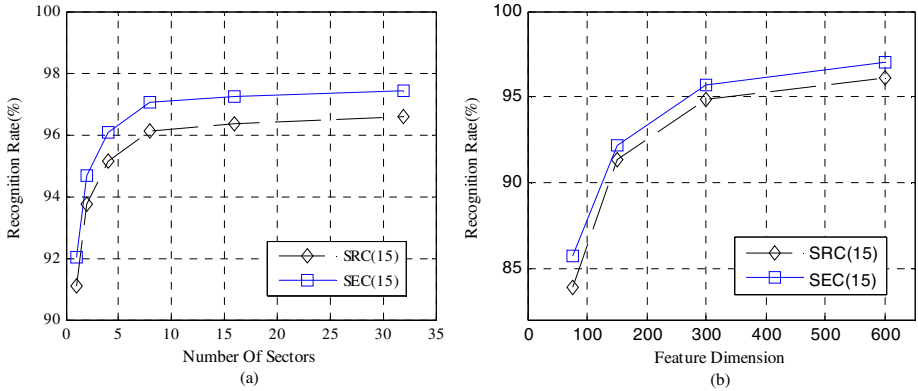


Fig. 1. Performance comparison. (a) Recognition effect for various number of sectors. (b) Recognition effect for various feature dimension.

Table 1. Comparisons of Recognition Rate on Iris Image Samples of Different Quality

Method	Image Quality		
	Good	Occluded	Seg. Incorrectly
SRC(10)	98.19	90.37	85.57
SEC(10)	98.74	91.12	86.73
SRC(15)	99.63	93.82	87.85
SEC(15)	99.89	94.96	88.53

6 Conclusion

In this paper, we introduced a SEC-based method for iris recognition, aiming to improve the robustness and efficiency of recognition process. A CSCI-based validation scheme relying on the coefficient vector got from the sparse signal representation was proposed herein. Under identical conditions, the SEC always outperforms the SRC method. With varied number of iris sectors, the training samples, and the feature dimension, our experimental results on CASIA-IrisV3 indicate the proposed method achieves excellent recognition performance both in robustness and efficiency when a sufficiently large number of training samples are available. Our future study will take into account improving the robustness of iris segmentation and developing the dictionary learning algorithm.

Acknowledgements. This work was supported by National Natural Science Foundation of China (Grant No. 61001123, 60973113), Science and Technology Planning Project of Hunan Province (No.2012FJ4129, 2012GK3056, S2013F1023), Science and Technology Planning Key Project of Changsha (No. K1104022-11, K1207027-11)

References

1. Donoho, D., Elad, M.: Maximal Sparsity Representation Via L_1 Minimization. *Proc. Nat. Aca. Sci.* 100, 2197–2202 (2003)
2. Wright, J., Yang, M., Ganesh, A., Sastry, S., Ma, Y.: Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)
3. Pillai, J.K., Patel, V.M., Chellappa, R., Ratha, N.K.: Secure And Robust Iris Recognition Using Random Projections And Sparse Representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(9), 1877–1893 (2011)
4. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(5500), 2323–2325 (2000)
5. Institute of Chinese Academy Of Sciences: CASIA Iris Image Database: (Version 3.0), <http://Biometrics.Idealtest.Org/Index.Jsp>
6. Chen, S., Donoho, D., Saunders, M.: Atomic Decomposition by Basis Pursuit. *SIAM Review* 43(1), 129–159 (2001)
7. Van Den Berg, E., Friedlander, M.P.: Probing the Pareto Frontier for Basis Pursuit Solutions. *SIAM Journal Scientific Computing* 31(2), 890–912 (2008)
8. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries For Sparse Representation. *IEEE Transactions on Signal Processing* 54(11), 4311–4322 (2006)

Face Verification across Age Progressing Based on Active Appearance Model and Gradient Orientation Pyramid

Xing Wu, Ji-Xiang Du, and Chuan-Min Zhai

Department of Computer Science and Technology, Huaqiao University, Xiamen 361021
wuxing_chris@163.com, {jxdu77, cmzhai}@gmail.com

Abstract. Face verification in the presence of age progression is an important problem that has not been widely addressed. In this paper, we propose to use the Active Appearance Model (AAM) and Gradient Orientation Pyramid (GOP) feature representation for this problem. We use the AAM on the dataset and then get the representation of Gradient Orientation on a hierarchical model, which the appearance of GOP. When combined with a support vector machine (SVM), the representation demonstrates excellent performance in our experiments.

Keywords: face verification, active appearance model, gradient orientation pyramid.

1 Introduction

Face verification is an important problem in computer vision and has a very wide range of applications, such as surveillance, human computer interaction, image retrieval, etc. There are many research effort has been focused on pursuing robustness to different imaging conditions, including illumination change, pose variation, expression, etc. Despite decades of study on face image analysis, age related facial image analysis has not been extensively studied until recently. Published approaches to age invariant face recognition are limited, Most of the available algorithms dealing with facial aging problem are focused on age estimation [1]-[5] and aging simulation [6]-[9]. Some previous work applied age progression for face verification tasks. When comparing two photos, these methods either transform one photo to have the same age as the other, or transform both to reduce the aging effects. Ramanathan and Chellappa [10] used a face growing model for face verification tasks for people under the age of eighteen. This assumption limits the application of these methods, since ages are often not available. Biswas et al. [11] studied feature drifting on face images at different ages and applied it to face verification tasks. Other studies used age transformation for verification includes [12]-[13]. A recent work in [14] used gradient orientation pyramid for feature representation, combined with support vector machine for verifying faces across age progression.

2 Active Appearance Model (AAM)

AAM (Active Appearance Model) is a feature point extraction method which widely used in the field of pattern recognition. Cootes first proposed the ASM [15], but the

ASM more or less ignores the texture (color and gray value) information of images. Then the AAM proposed, the facial features localization method based on AAM in establishing face model process not only considers the local features, but also comprehensive considers the global shape and texture information. Through statistical analysis of the shape and texture features of the human faces to establish face mixed model, which is the final AAM.

The model parameter c of the AAM used to control the shape and texture of the model. The model shape and texture presentation:

$$\begin{aligned} x &= \hat{x} + Q_s c \\ g &= \hat{g} + Q_g c \end{aligned} \quad (1)$$

Where \hat{x} is the average vector of model shape, \hat{g} is the average vector of model texture, Q_s and Q_g are the presentation matrix of the change of the descriptive model after training the samples.

Then we use a linking vector to link the shape and texture of a sample:

$$b = \begin{pmatrix} W_s s \\ g \end{pmatrix} = \begin{pmatrix} W_s Q_s c \\ Q_g c \end{pmatrix} \quad (2)$$

Where W_s is the linking vector which is a diagonally matrix. Then we use PCA on this linking vector, the model we get:

$$b = \hat{b} + Q_b c \quad (3)$$

Where \hat{b} is the average vector of the linking vectors?

After we get the AAM by this method, we use the 68 feature points which we labeled manually to normalize the AAM. The finally AAM we used are shown in the Fig.1.

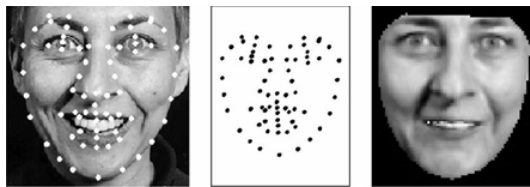


Fig. 1. The AAM image and the normalized image

3 Gradient Orientation Pyramid (GOP)

Because of previous study of the robustness of gradient orientation (GOP) [14], we propose to use it as the feature descriptor for face verification across age progression in our experiment. After we get the AAM of all the images, given an AAM image $I(p)$, where $p = (x, y)$ indicates pixel locations, we first define the pyramid of I as $P(I) = \{I(p; \sigma)\}_{\sigma=0}^s$ with:

$$\begin{aligned}
I(p;0) &= I(p) \text{ ,} \\
I(p;\sigma) &= [I(p;\sigma-1) * \Phi(p)] \downarrow_2 \text{ ,} \quad \sigma = 1, \dots, s
\end{aligned}
\tag{4}$$

Where $\Phi(p)$ is the Gaussian kernel (0.5 is used as the standard deviation in our experiment), \otimes denotes the convolution operator, \downarrow_2 denotes half size down-sampling, and s is the number of pyramid layers. Note that in (4) the notation I is used both for the original image and the images at different scales for convenience.

Then, the gradient orientation at each scale σ is defined by its normalized gradient vector at each pixel:

$$g(I(p;\sigma)) = \begin{cases} \frac{\nabla(I(p,\sigma))}{|\nabla(I(p,\sigma))|} \text{ ,} & \text{if } |\nabla(I(p,\sigma))| > \tau \\ (0,0)^T \text{ ,} & \text{otherwise} \end{cases}
\tag{5}$$

Where τ is a threshold for dealing with “flat” pixels. The gradient orientation pyramid (GOP) of I , is naturally defined as $\varphi(I) = \text{stack}(\{g(I(p,\sigma))\}_{\sigma=0}^s) \in R^{d \times 2}$ that maps I to a $d \times 2$ representation, where $\text{stack}(\cdot)$ is used for stacking gradient orientations of all pixels across all scales and d is the total number of pixels.

Given an AAM image pair (I_1, I_2) and corresponding GOPs, the feature vector $x = F(I_1, I_2)$ is computed as the cosines of the difference between gradient orientations at all pixels over scales:

$$x = F(I_1, I_2) = (G_1 \odot G_2) \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\tag{6}$$

Where \odot is the element-wise product. Next, we apply the Gaussian kernel to the extracted feature x to be used with the SVM framework. Specifically, our kernel is defined as:

$$K(x_1, x_2) = \exp(-\gamma |x_1 - x_2|^2)
\tag{7}$$

Where γ is a parameter determining the size of RBF kernels.

4 Experiments

In this section, we report experimental results obtained on MORPH database by comparing our algorithm with a number of related face verification methods.

4.1 Experimental Datasets

In this section, we report results on experiments on a large public domain face aging data set, which is the MORPH database. While there are several public domain face datasets, only a few are constructed specifically for the aging problem. The lack of a large face aging database until recently limited the research on age invariant face recognition.

In the MORPH database, there are totally 52099 facial images of 12938 subjects in the age range 16-77 years. After get the AAM of an image, which the size is 120*126 pixels, and then we normalize the AAM image into 70*80 pixels and transform it into a gray-scale one.



Fig. 2. Example images of MORPH database



Fig. 3. The images of AAM



Fig. 4. The nominalized and gray-scale images of AAM

4.2 Experimental Compared Approaches

We compare the following approaches: 1) SVM+AAM+GOP: the approach proposed in this paper. 2) SVM+AAM+GO: this is the approach proposed without a hierarchical model. 3) SVM+GOP: this is the method using GOP in [14]. 4) SVM+GO: SVM+GOP without a hierarchical model. 5) SVM+G: this one is similar to SVM+GO, except that the gradient (G) is used instead of gradient .6) SVM+diff: [16]. As in [16], we use the differences of normalized images as input features combines with SVM. 7) GO: this is the method using gradient orientation proposed in [17]. 8) L_2 : this is a baseline approach that uses the L_2 norm to compare two normalized images.

4.3 Experimental Evaluation

For verification tasks, the correct reject rate (CRR) and the correct acceptance rate (CAR) are two critical criteria:

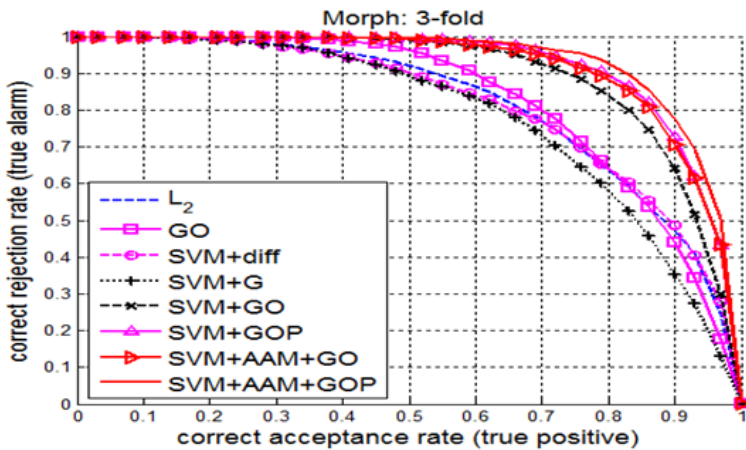
$$CRR = \frac{\# \text{correctly rejected extra-personal pairs}}{\# \text{total extra-personal pairs}}, \quad (8)$$

$$CAR = \frac{\# \text{correctly accepted intra-personal pairs}}{\# \text{total extra-personal pairs}},$$

Where “accept” indicates that the input image pair are from the same subject and “reject” indicates the opposite. In addition, the equal error rate (EER), defined as the error rate when a solution has the same CAR and CRR, is frequently used to measure verification performance.

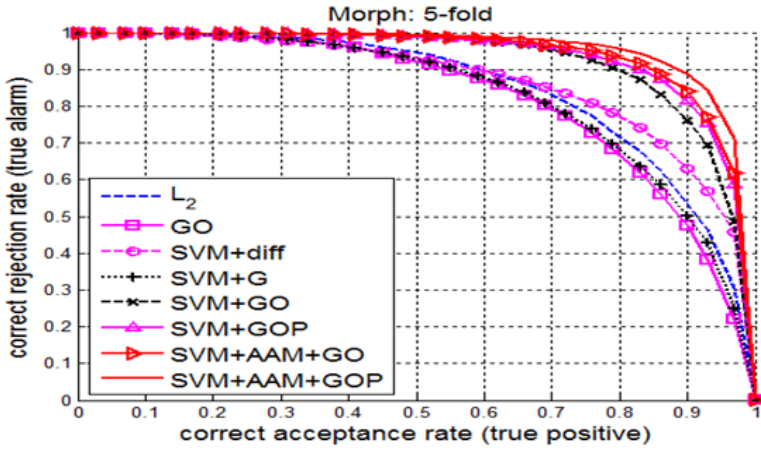
4.4 Experimental Results

For verification tasks, three-fold cross validation is used, such that in each fold samples from the same subject never appear in both training and testing pairs. Each fold contains 2000 intra-personal pairs and 2000 extra-personal pairs which we choose randomly from all the intra-personal pairs and extra-personal pairs by collecting all image pairs from all the subjects. In addition, we also use the five-fold and ten-fold cross validation to improve the verification performance of our method. The experiment results are shown in Fig.5 and Table 1.

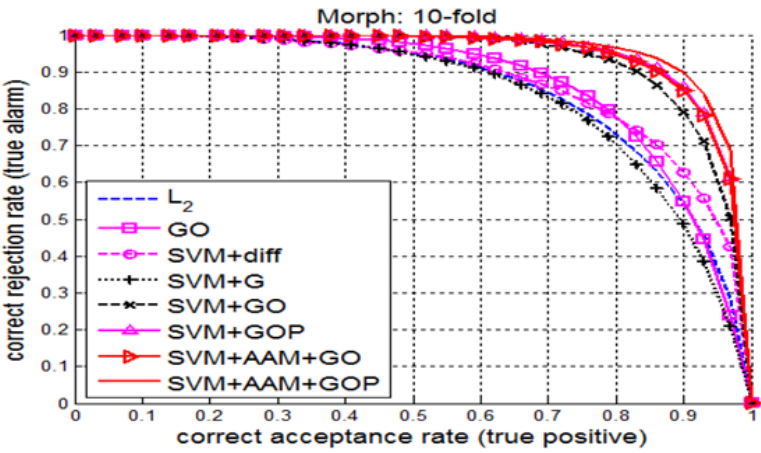


(a)

Fig. 5. CRR-CAR curves for cross validation experiment on MORPH dataset(a) three-fold (b) five-fold (c) ten-fold



(b)



(c)

Fig. 5. (Continued)

Table 1. Equal error rates for experiment on the MORPH database

EER	3-fold	5-fold	10-fold
L2	0.26647	0.23506	0.22928
GO	0.25574	0.25514	0.20713
SVM+diff	0.26848	0.21410	0.21112
SVM+G	0.28612	0.25039	0.23595
SVM+GO	0.18358	0.15204	0.13755
SVM+GOP	0.15570	0.13297	0.11583
SVM+AAM+GO	0.15949	0.12567	0.11960
SVM+AAM+GOP	0.14271	0.10677	0.10049

From the figures and table, we study that the method proposed outperforms all the others. In addition, the EER decreases with the n of the n -fold cross validation increases.

5 Conclusion

In this paper we studied the problem of face verification with age variation combining Active Appearance Model (AAM) and Gradient Orientation Pyramid (GOP) representation. First, we establish the AAM on the MORPH database. After getting the robust descriptor, the gradient orientation pyramid, we use SVM classification to train and test it. To compare to previously used descriptors, the proposed method performs well on face images with large age differences.

We plan to test on others large public datasets for deeper understanding of the proposed approaches. In addition, we anticipate the proposed and more effective feature extraction methods to solve face verification problem in the future research.

Acknowledgments. This work was supported by the grants of the National Science Foundation of China (Nos. 61175121&61102163), the Program for New Century Excellent Talents in University (No.NCET-10-0117), the grants of the National Science Foundation of Fujian Province (No.2011J01349, 11136006), the Program for Excellent Youth Talents in University of Fujian Province (No.JA10006), the Fundamental Research Funds for the Central Universities of Huaqiao University (No.JB-SJ1003, 09HZR15).

References

1. Fu, Y., Huang, T.S.: Human Age Estimation with Regression on Discriminative Aging Manifold. *IEEE Tran. on Multimedia* 10(4), 578–584 (2008)
2. Geng, X., Zhou, Z., Smith-Miles, K.: Automatic Age Estimation Based on Facial Aging Patterns. *IEEE Tran. on Pattern Analysis and Machine Intelligence* 29(12), 2234–2240 (2007)
3. Guo, G., Fu, Y., Dyer, C., Huang, T.: Image-Based Human Age Estimation by Manifold Learning And Locally Adjusted Robust Regression. *IEEE Tran. on Image Processing*. 17(7), 1178–1188 (2008)
4. Kwon, Y.H., Lobo, N.D.V.: Age Classification from Facial Images. *Computer Vision and Image Understanding* 74(1), 1–21 (1999)
5. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing Different Classifiers for Automatic Age Estimation. *IEEE Tran. on Systems, Man, and Cybernetics, Part B: Cybernetics* 34(1), 621–628 (2004)
6. Lanitis, A., Taylor, C., Cootes, T.: Toward Automatic Simulation of aging Effects on Face Images. *IEEE Tran. on Pattern Analysis and Machine Intelligence*. 24(4), 442–455 (2002)
7. Suo, J., Zhu, S., Shan, S., Chen, X.: A Compositional And Dynamic Model for Face Aging. *IEEE Tran. on Pattern Analysis and Machine Intelligence*. 32(3), 385–401 (2010)

8. Tsumura, N., Ojima, N., Sato, K., Shiraishi, M., Shimizu, H., Nabeshima, H., Akazaki, S., Hori, K., Miyake, Y.: Image-Based Skin Color and Texture Analysis/Synthesis by Extracting Hemoglobin and Melanin Information in the Skin. *ACM Transactions on Graphics (TOG)* 22(3), 770–779 (2003)
9. Park, U., Tong, Y., Jain, A.K.: Age Invariant Face Recognition. *IEEE Tran. on Pattern Analysis and Machine Intelligence* 32(5), 947–954 (2010)
10. Ramanathan, N., Chellappa, R.: Face Verification across Age Progression. *IEEE Tran. on Image Processing* 15(11), 3349–3361 (2006)
11. Biswas, S., Aggarwal, G., Ramanathan, N., Chellappa, R.: A Non-Generative Approach for Face Recognition across Aging. In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2008, pp. 1–6. IEEE Press, Arlington (2008)
12. Patterson, E., Sethuram, A., Albert, M., Ricanek, J.K., King, M.: Aspects of Age Variation on Facial Morphology Affecting Biometrics. In: First IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2007, pp. 1–6. IEEE Press, Crystal City (2007)
13. Singh, R., Vatsa, M., Noore, A., Singh, S.K.: Age Transformation for Improving Face Recognition Performance. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) *PRMI 2007*. LNCS, vol. 4815, pp. 576–583. Springer, Heidelberg (2007)
14. Ling, H., Soatto, S., Ramanathan, N., Jacobs, D.: Face Verification across Age Progression Using Discriminative Methods. *IEEE Tran. on Information Forensics and Security* 5(1), 82–91 (2010)
15. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. *IEEE Tran. on Pattern Analysis and Machine Intelligence* 23(6), 681–685 (2001)
16. Phillips, P.J.: Support Vector Machines Applied to Face Recognition. In: *Advances in Neural Information Processing Systems 16 (NIPS)*, vol. 2, pp. 803–809. MIT Press, Cambridge (1999)
17. Chen, H., Belhumeur, P., Jacobs, D.: In Search of Illumination Invariants. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 254–261. IEEE Press, Hilton Head Island (2000)

Single Sample Face Recognition Based on DCT and Local Gabor Binary Pattern Histogram

Zhihua Xie

Key Lab of Optic-Electronic and Communication,
Jiangxi Sciences and Technology Normal University, Nanchang, Jiangxi, 330013, China
xie_zhihua@aliyun.com

Abstract. For years, researchers in face recognition area have been representing and recognizing faces based on subspace discriminant analysis or statistical learning. Nevertheless, for single sample face recognition these approaches are always suffering from the generalizability problem because of small samples. This paper proposes a novel non-statistics features extraction approach based on fusion of DCT and local Gabor binary pattern Histogram (LGBPH). The global and low frequency features are obtained by low frequency coefficients of discrete cosine transform (DCT). The local and high frequency features are extracted by LGBPH. To integrate the global and local features, the final recognition can be achieved by parallel integration of classification results of the global and local features. In DCT and LGBPH, training procedure is unnecessary to construct the face model, so that the generalizability problem is naturally avoided. The experimental results on ORL face databases show that the global face and local information can be integrated well after level fusion by global and local features, which improve the performance of single sample face recognition.

Keywords: global feature, local feature, DCT, LGBPH, non-statistics feature extraction.

1 Introduction

As one of the few biometric methods that possess the merits of both high accuracy and low intrusiveness, face recognition has a variety of potential applications in information security, law enforcement and surveillance, smart cards, access control, among others [1]. The area of face recognition is well-described today in many papers and books [2, 3]. The idea that two-dimensional still-image face recognition in controlled environment is already a solved task is generally accepted and several benchmarks evaluating recognition results were done in this area (e.g. Face Recognition Vendor Tests, FRVT 2000, 2002, 2006) [7]. Nevertheless, many tasks have to be solved, such as recognition in unconstrained environment, recognition of non-frontal images, single sample per person problem, etc [6]. As a great challenge, the single sample per person problem has become a big obstacle to many real-world applications, such as e-passport, watch list screening, since, in these scenarios, it is generally impossible to collect more than one sample per person.

Numerous methods have been proposed to improve face recognition accuracy. These methods can be roughly divided into two categories: geometry based methods and appearance-based methods [7, 11]. The former describe a face using the relationship of facial components, e.g., the relative position of eyes and nose, while the latter represent the face using holistic appearance. In recent years, appearance-based methods have become the dominant approaches for face recognition. Many appearance-based approaches have been proposed in the previous work, nevertheless most of them are based on statistical learning, such as principal component (PCA) [4], linear discriminant analysis (LDA) [4], locality preserving projection (LPP) [5], which can inherently be heavily affected by the number of training samples for each person. Therefore, the representation methods based on statistical learning are not suitable for single sample face recognition [9]. Appearance-based methods using image content itself are appreciated for single sample face recognition [3, 7].

This paper focuses on the useful feature extraction based on image content for single sample face recognition. To extract the global features in face image, discrete cosine transform (DCT) is applied to preserve the low frequency coefficients. Local Gabor binary pattern histogram (LGBP) [3] can effectively provide local structure information and high frequency components. The main idea of this paper is to combine DCT and LGBP features to extract not only global and local but also low frequency and high frequency information for single sample face recognition. Meanwhile, the proposed approach is computationally simple and is training-free: there is no need to learn a dictionary and no tuning of parameters.

2 Face Representation Based on Discrete Cosine Transform

The traditional extraction methods are based on statistical character, such as PCA (principle component analysis), ICA (independent component analysis), LDA (linear discriminative analysis) etc [4, 5]. Those methods suffer from the problems of complexity of computation, large memory units and small samples. To tackle with those problems, discrete cosine transform (DCT) is chosen to extract the global features of the face image. As we known, a large amount of information about the original image is stored in a fairly small number of coefficients (in the upper-left corner, corresponding to low spatial frequency DCT components in the image) [12]. This phenomenon can be used for character extraction. Another merit of the DCT is it can be implemented efficiently using the Fast Fourier Transform (FFT). So DCT can be widely used in real-time face recognition system. For a digital image $f(x, y)$ with the resolution $M \times N$, its 2-dimensional DCT is defined as [12]:

$$A(i, j) = a(i)a(j) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \times \cos \left[\frac{(2x+1)i\pi}{2M} \right] \cos \left[\frac{(2y+1)j\pi}{2N} \right] \quad (1)$$

$$i = 0, 1, 2, \dots, M-1; j = 0, 1, \dots, N-1$$

where $A(i, j)$ is the result of DCT, namely, the coefficient of DCT. $a(i)$, $a(j)$ are defined as:

$$a(i) = \begin{cases} \sqrt{1/M}, & i = 0 \\ \sqrt{2/M}, & i = 1, 2, 3, \dots, M - 1 \end{cases} \quad (2)$$

$$a(j) = \begin{cases} \sqrt{1/N}, & j = 0 \\ \sqrt{2/N}, & j = 1, 2, 3, \dots, N - 1 \end{cases}$$

The characteristic of DCT coefficients distribution is that most of non-zero DCT coefficients locate at the low-frequency region (the top left corner, where both i and j are small). This region contains most useful information, energy, and features.

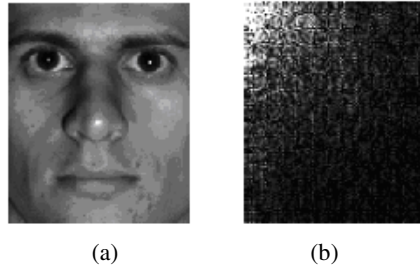


Fig. 1. (a) face image (b) its DCT coefficients

In this paper, we select the low-frequency DCT coefficients to extract the global features of face image, which can reduce the dimensions of the face image.

3 Face Representation Based on LGBP

The LGBP is proposed by Zhang et al [3]. The Gabor wavelets have been found to be particularly suitable for image decomposition and representation when the goal is the derivation of local and discriminating features [3]. The multi-resolution and multi-orientation Gabor filters can extract different textures and local information in face. LBP is an effective method to represent the local and high-frequency face information. The overall framework of the proposed representation approached based on Local Gabor Binary Pattern Histogram sequence (LGBPHs) [3] is illustrated in Fig.2. In this approach, a face image is modeled as a “histogram sequence” by the following procedure: (1) An input face image is normalized and transformed to obtain multiple Gabor Magnitude Pictures (GMPs) in frequency domain by applying multi-scale and multiorientation Gabor filters; (2) Each GMP is converted to Local Gabor Binary Pattern (LGBP) map; (3) Each LGBP Map is further divided into non-overlapping rectangle regions with specific size, and histogram is computed for each region; (4) The LGBP histograms of all the LGBP Maps are concatenated to form the final histogram sequence as the model of the face.

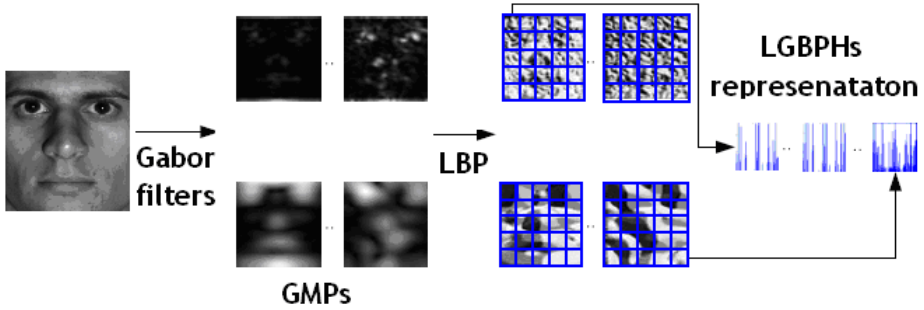


Fig. 2. Framework of LGBPHs

In this paper, representation based on LGBPHs is used as local and high-frequency features of facial image. We choose Gabor filter banks with 5 scales and 8 orientations and LBP with $r=1$ and $p=8$ for LGBPH sequence.

4 Fusion of DCT and LGBPH

As we know, a combination of many different classifiers can improve predictions if there is a big difference between each classifier [8]. Classifier based on Local features and Classifier based on global features can contribute complementary discriminant ability for face recognition [10]. In this paper, we integrate global classifier and local classifier by weighted summation in parallel manner. The flowchart of fusion is shown in Fig 3.

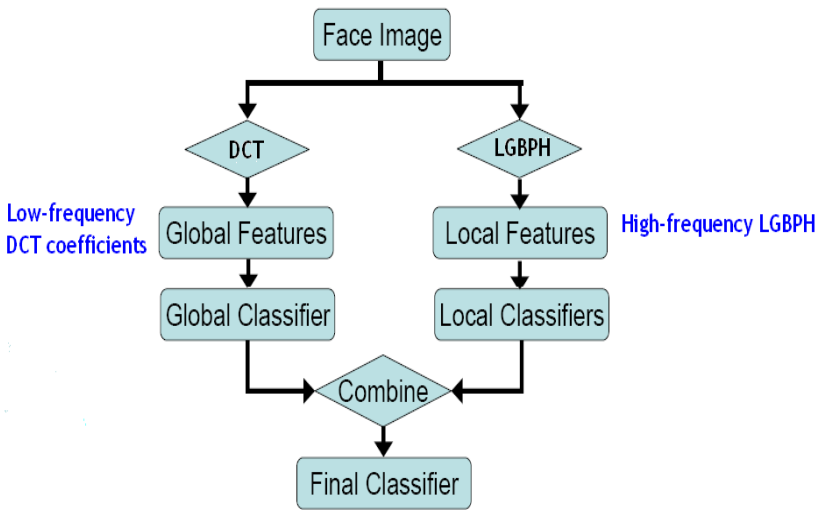


Fig. 3. Flowchart of fusion

For describing conveniently, the global feature and the local feature vector are named global classifier (referred to as GC) and local classifier (referred to as LC) respectively. The final classifier (referred to as FC) can be gotten by parallel integration of classification results of the DCT and LGBPH features:

$$FC = w_G \cdot GC + w_L \cdot LC \quad (3)$$

Where w_G is the contribution rate of feature based on DCT, $w_L = 1 - w_G$ is the contribution rate of features based on LGBPHs.

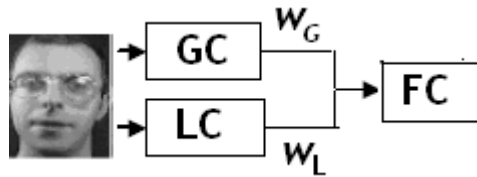


Fig. 4. Construction of global classifier, local classifier and final classifier

Since the features based on DCT and features based on LGBPHs have different benchmarks, it is reasonable to fuse local and global features based on the same normalized classifier. In this paper, an appropriate membership grade is defined to results of GC and LC, which ranges from 0 to 1. If the test sample is i , its membership grade belong to j th class can be expressed by v_j :

$$v_j = 1/[1 + (d_{ij} / D)] \quad (4)$$

Where d_{ij} is the distance between test sample and j th class, D is the sum of distance between test sample and all training samples.

$$D = \sum_j d_{ij} \quad (5)$$

5 Experiment Results

In order to evaluate the performances of the proposed face recognition method, the ORL is used in the experiments [12]. This database contains photos from faces of 40 people, each person in 10 different poses, totalizing 400 photos. The faces were photographed at different moments, with varying lighting, facial expressions (eyes closed/opened, smiling/not smiling), facial poses and facial details (with/without glasses, with/without beard), among other types of variations. The images are in grayscale, with 256 different levels and a dimension of 92x112 pixels. Some examples are shown in Fig 5. In our experiments, the first photo of each person is used to construct training samples database and the remainder photos (360 photos) are for testing.



Fig. 5. Examples from the ORL database: persons 19, 11 and 36 at their pose 1 to 5

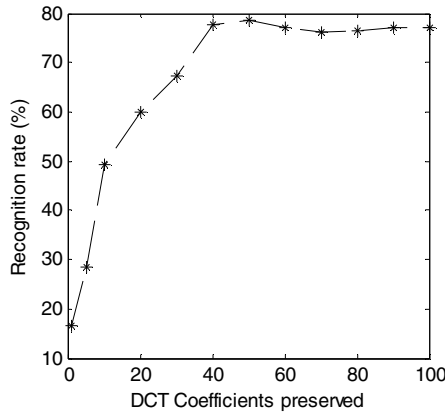


Fig. 6. Recognition rates versus different number of DCT coefficients

To evaluate the performance of face representation based on DCT, “Z” scanning is applied to preserve low-frequency DCT coefficients. Different low-frequency coefficients are preserved for face recognition, the results are shown in Fig.6. From Fig. 6, it is clear that only a small part of DCT coefficients has the contribution to face recognition.

In features extraction based on LGBPHs, the variable parameter is the mode of partitioning. Five modes of partitioning are used: non- partitioning, 2×2, 4×2, 2×4 and 4×4. Recognition results with different partitioning modes are shown in Table 1.

Table 1. Results of different partitioning modes

Partitioning modes	1×1	2×2	2×4	4×2	4×4
Recognition rate %	71.4	83.33	86.11	85.00	86.67

It can be seen from the Table 1 that the feature extraction method based on LGBPH get can the best recognition rate at partitioning modes 4×4 and partitioning mode can improve the recognition performance.

To make full use of discriminant ability in different features, the local features and global features have different contributions to final face classifier. Fig.7 shows face recognition results with different weights of GC. As seen, the performance of final classifier is optimal when w_G is about 0.2. It is significant to point out that the weight of LC is larger than that of GC. This means discriminant ability of local features is higher than that of global features. This experiment result also proved our point of view before: global and local feature play a different role in face representation, including complementary discriminant information, therefore, both of them should be integrated.

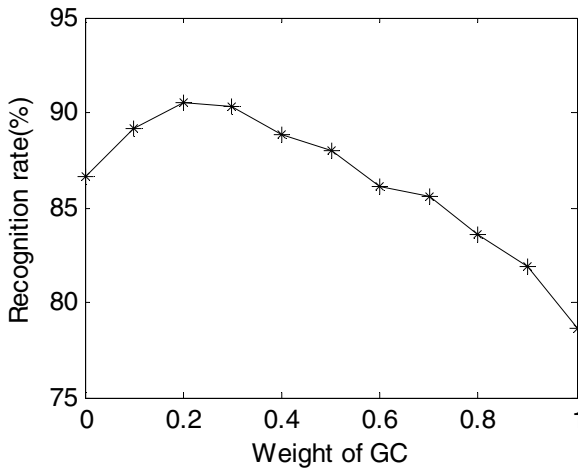


Fig. 7. Recognition results with different w_G

The best recognition results based on different methods are shown in Table 2. It can be seen from Table 2 that the DCT +LGBPH method for single sample face recognition can get better recognition performance compared with DCT+LDA and LGBPH methods.

Table 2. best Results of different methods

Methods	Recognition rate (%)
DCT+LDA [12]	78.61
LGBP[3]	86.67
DCT+LGBP	90.56

6 Conclusion

This paper focuses on features extraction based on non-statistical learning for single sample face recognition. A novel formulation based on integration of global and local

feature is proposed to improve the performance of single sample face recognition. Compared to previously method such as traditional DCT and LGBPHs, the new method is more robust and performs well in single sample face recognition. In the future, we need to research more how the global and local feature interact each other in human visual perception system and it will be a challengeable research task.

Acknowledgments. This paper is supported by National Nature Science Foundation of China (No. 61201456) and Natural Science Foundation of Jiangxi Province of China (No.20132BAB201052)

References

1. Tan, X., Chen, S., Zhou, Z., And Zhang, F.: Face Recognition from A Single Image per Person: A Survey. *Pattern Recognition* 39(9), 1725–1745 (2006)
2. Zhao, W.Y., Chellappa, R., Phillips, P.J., Rosenfeld, A.P.: Face Recognition: a literature survey. *ACM Computing Surveys* 35(4), 399–458 (2003)
3. Zhang, W., Shan, S., Gao, W., et al.: Local Gabor Binary Pattern Histogram Sequence (LGBPHS):A Novel Non-Statistical Model for Face Representation and Recognition. In: *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV 2005)*, pp. 786–791 (2005)
4. Loog, M., Duin, R., Haeb-Umbach, R.: Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(7), 762–766 (2001)
5. Hamsici, O.C., Martinez, A.M.: Bayes Optimality in Linear Discriminant Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4), 647–657 (2008)
6. Kao, W., Hsu, M., Yang, Y.: Local Contrast Enhancement And Adaptive Feature Extraction for Illumination-Invariant Face Recognition. *Pattern Recognition* 43(5), 1736–1747 (2010)
7. Kan, M., Shan, S., Su, Y., Chen, X., Gao, W.: Adaptive Discriminant Analysis for Face Recognition from Single Sample Per Person. In: *FG 2011*, pp. 193–199 (2011)
8. Yu, S., Shan, S., Chen, X., et al.: Integration of Global and Local Feature for Face Recognition. *Journal of Software* 21(8), 1849–1862 (2010)
9. Wang, J., Plataniotis, K.N., Lu, J., et al.: On Solving The Face Recognition Problem with One Training Sample per Subject. *Pattern Recognition* 39(9), 174–1762 (2006)
10. Kou, J., Du, J.-X., Zhai, C.-M.: Integration of Global And Local Feature for Age Estimation of Facial Images. In: Huang, D.-S., Ma, J., Jo, K.-H., Gromiha, M.M. (eds.) *ICIC 2012*. LNCS, vol. 7390, pp. 455–462. Springer, Heidelberg (2012)
11. Lu, J., Tan, Y., Wang, G.: Discriminative Multi-Manifold Analysis for Face Recognition from A Single Training Sample per Person. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1), 39–51 (2013)
12. Zhang, Y., Liu, C.: A Novel Face Recognition Method Based on Linear Discriminant Analysis. *Journal of Infrared and Millimeter Waves* 22(5), 327–330 (2003)

Iris Classification Based on Its Quality

Aditya Nigam, Anvesh T, and Phalguni Gupta

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur,
Kanpur-208016, India
{naditya, anvesh, pg}@cse.iitk.ac.in

Abstract. This paper proposes an iris classification method based on the iris image quality. Quality of an iris image is modeled as a function of the attributes like focus, motion blur, occlusion, contrast and illumination, specular reflection and dilation. Values of these attributes are combined using a support vector machine (SVM) to provide the overall quality class of the image.

Keywords: SVM, Image Quality, Iris, Biometrics, Focus.

1 Introduction

The use of biometrics for recognition of individuals is growing rapidly. Behavioral as well as physiological biometrics based characteristics (such as face [13, 14, 10], fingerprint [17], iris [4,15], palmprint [3,18], knuckleprint [12,2], gait, voice, vein patterns etc.) are used to develop robust, accurate and highly efficient personal authentication systems. Among all of the biometric traits, iris is the most reliable one, both in terms of recognition and identification performance because iris texture has high variation among inter class samples. Quality of the image plays an important role in any identification system because higher quality images have a lower false accept and false reject rates. The quality of the iris image is affected by the presence of eyelashes and eyelids on the iris region, lack of focus on the iris region, improper lightning condition, dilation of the iris etc.

Several attempts have been made to estimate the quality of iris images. In [6] the focus of the iris image can be estimated by calculating the total high frequency power in the 2D Fourier spectrum of the image. In [20] iris quality estimation is done by combining three image quality attributes, viz occlusion, focus and motion blur. It computes the high frequency output of 2D Fourier spectrum of the image using a 5×5 kernel which is convolved with the image and the output is compared against a predefined threshold to determine whether the image is out of focus or not. The use of Dempster-Shafer theory [16] for iris image quality has been proposed in [9]. Various quality attributes like defocus blur, motion blur, occlusion, off-angle, specular reflection, lightning and pixel count are considered to obtain the final quality score. In [5] 2D continuous wavelet is applied on the concentric band of the iris image. In [19] Laplacian of Gaussian operator (LoG) has been proposed for iris quality estimation that does not rely on the localization of the iris.

In this paper, a method is proposed to classify iris images into different quality classes lying between 1 and 5. Higher class indicates higher quality of the iris image. The paper is organized as follows: Section 2 discusses the proposed method of iris quality estimation. Section 3 contains the experimental results followed by the concluding section.

2 Iris Classification Based on Its Quality

The quality of the iris image is modeled as a function of the following six attributes: Focus (F), Motion Blur (MB), Occlusion (O), Contrast and Illumination (CI), Dilation (D), Specular Reflection (SR). Hence overall quality of the iris image can be represented as:

$$Quality = f(F, MB, O, CI, D, SR) \quad (1)$$

where f is a function which is learned from the training data using a Support Vector Machine (SVM). Quality attributes from the training set of iris images along with their respective true quality label are used to train the SVM classifier. This creates a model based classifier which is later used to predict the quality of a query iris image using its quality attributes.

Iris Segmentation. The input iris image is segmented to identify the iris region in the image using the procedure described in [4]. Canny edge detection is used to detect edges on the thresholded image and an improved circular hough transform is used to detect the pupil boundary. A window of $w \times w$ around the center of pupil is used as a possible search space for center of the iris and its outer boundary is identified using integro-differential operator as defined in [7]. The annular iris region is converted to polar coordinate system to obtain a rectangular strip that can deal with dimensional inconsistencies between different human eye images. This normalization process produces iris region of constant dimensions.

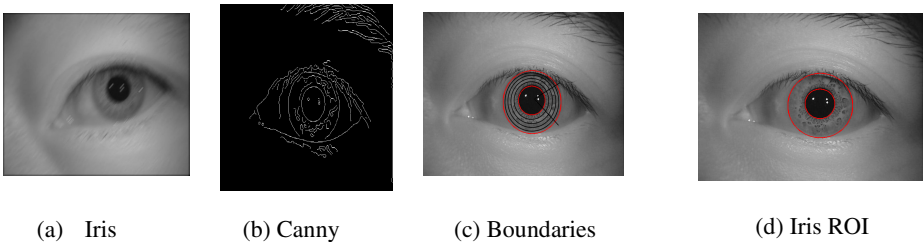


Fig. 1. Iris Segmentation Steps

2.1 Proposed Quality Attributes

The segmented iris image is used to calculate the image quality attributes individually. These attributes are passed to the Support Vector Machine which

classifies the iris image into an appropriate class. The calculation for each of the six image quality attributes is given as follows.

[1] Focus (F) : Focus of an image refers to the amount of defocus blur in the image. Defocus blur is the blurring introduced in the image when the focal point is outside the depth of the field of object being captured. Greater the distance, greater is the amount of defocus blur. The 2D Fourier spectrum of any well focused image is uniform but for a defocus image the spectrum is concentrated more towards the low frequencies. The focus of the image can be estimated by calculating the high frequencies in the image spectrum. This is done by convolving the image with a newly proposed 6×6 kernel. This kernel well approximates the 2D Fourier spectrum high frequency band pass filter. The higher the response from the image, the more focused it is. The proposed 6×6 kernel is given below and its filter responses is shown in Figure 2.

$$K = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 8 & 8 & -1 & -1 \\ -1 & -1 & 8 & 8 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix} \quad (2)$$

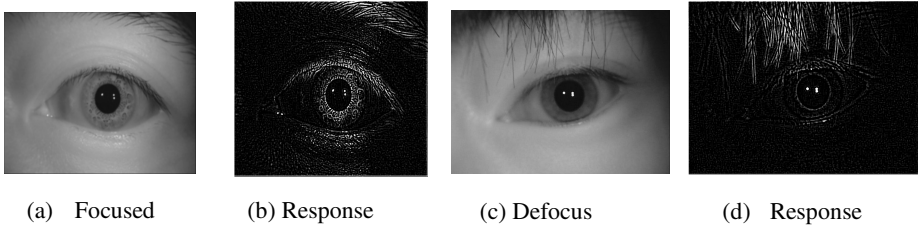


Fig. 2. Defocus Filter Response

[2] Motion Blur (MB) : It is the blurring effect caused due to the movement of the camera relative to the subject or vice-versa during image acquisition. It is characterized by the degree of shear and the direction of the blur. The degree of shear indicates the amount of distortion in the image along the direction of the motion blur. The estimation of the motion blur in the image is done using Just Noticeable Blur (JNB) [8] which is the least amount of blur at which the blur in the edge is noticeable. The edge width at which JNB occurs is called JNB width of the edge. The image is divided into blocks of equal size. A block is processed if it contains edge pixels above

a certain threshold. The JNB value is calculated as a cumulative blur of all blocks in the image. The blur metric is calculated for all edges in a block of an image. In [8] the probability of blur detection is given by the psychometric function

$$P(e_i) = 1 - e^{-\left(\frac{w(e_i)}{w_{jnb}(e_i)}\right)^\beta} \quad (3)$$

where $P(e_i)$ is the probability of blur detection and $w(e_i)$ and $w_{jnb}(e_i)$ are the measured width and the JNB width of an edge. This function is used to calculate the probability of blur in a block B by using

$$P(B) = 1 - \prod_{e_i \in B} (1 - P(e_i)) \quad (4)$$

The perceived blur in the block B, is denoted by D_B and is calculated as

$$D_B = \left(\sum_{e_i \in B} \left| \frac{w(e_i)}{w_{jnb}(e_i)} \right|^\beta \right)^{\frac{1}{\beta}} \quad (5)$$

where β is the median parameter obtained based on the JNB values. The overall blur distortion measure is calculated in [8] for all blocks as

$$D = \left(\sum_R |D_B|^\beta \right)^{\frac{1}{\beta}} \quad (6)$$

The blur in an edge is detected if the measured width of the edge is at least equal to the JNB width of the edge. JNB width of the edge is calculated based on the local contrast of the region. When the JNB width is equal to the measured edge width, the probability of detection of the edge is 63% (from Eq. 3). As the edge width increases, the probability of detection of blur increases.

[3]Occlusion (O) : Occlusion in an iris image is caused due to the presence of eyelids or eyelashes on the iris region. This is the most common and significant image quality degradation attribute. The occlusion is calculated as

$$Occlusion = \frac{\text{Number of non - occluded pixels}}{\text{Total number of pixels in the image}} \quad (7)$$

The number of pixels that are occluded can be calculated by the sum of the number of pixels occluded by the eyelids and by the eyelashes. The image is smoothed using a Gaussian filter to reduce noise in the image. Canny edge detection is performed on the smoothed image to detect the eyelid edges. The edge map image is dilated so that any edges which are broken during edge detection are joined. The dilated image contains

the eyelid edges which are used to identify the distinct bounded regions in the iris image using a connected component analysis on the image. Eyelashes can be identified by their low pixel values. The input image is thresholded using an empirically selected threshold to obtain the position of all the eyelashes in the form of a binary image. The eyelash mask is combined (i.e logical OR) with the eyelid mask to obtain the full occlusion mask. Various stages of occlusion detection are shown in Figure 3.

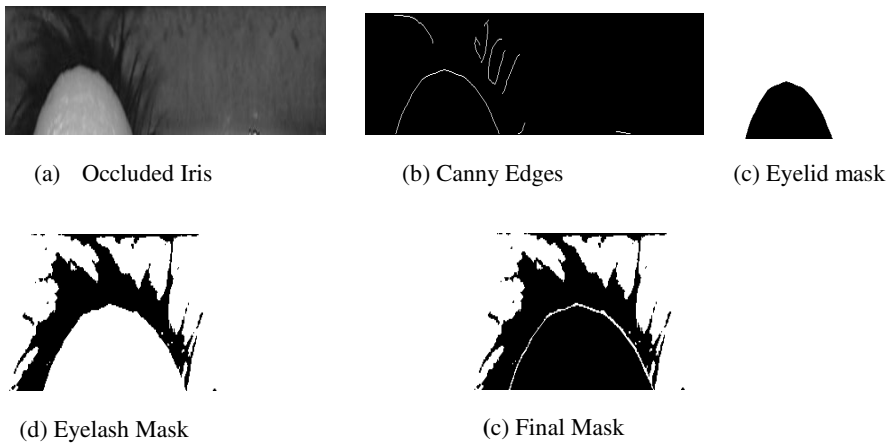


Fig. 3. Iris Occlusion Mask

[4] Contrast and Illumination (CI): Contrast of an image is defined as the range of its intensity levels. Any high contrast image generally has a uniformly distributed histogram. This parameter tries to identify the images that are either too dark or too light. It predicts whether the image is uniformly illuminated or not. The contrast is estimated by ignoring the extreme values which indicates the image is too bright or too dark. The range of pixel intensities is divided into three groups are empirically selected as (0,35), (36,220), (221,255). The intensities in the region (36,220) indicate moderate intensity levels. The ratio of the number of the pixels in this region to the total number of pixels is used to quantify the uniformity in illumination throughout the image.

[5]Dilation (D): Dilation of the pupil is a natural process of expansion of the pupil area due to the lack of proper lightning within the surroundings that reduces the amount of iris area. The dilation metric is defined as the ratio of the iris region available to the total area of the iris outer boundary circle. The dilation of the image affects severely the normalization of the iris image. Higher the dilation in the iris images, more is the pixel value redundancy in the normalized image. This also causes the iris texture pattern to shift its spatial position within the normalized iris image. Thus the dilation attribute can be estimated as :

$$dilation = \frac{\text{Area of iris}}{\text{Area of iris outer circle}} = 1 - \left(\frac{r_p^2}{r_i^2} \right) \quad (8)$$

where r_i and r_p are the radius of the iris outer and pupil's outer boundary respectively and are obtained during iris segmentation step [4].

[6] Specular Reflection (SR): Specular reflections are caused due to light source that is used to illuminate the iris. Iris is very reflective, so the light reflection is visible in the acquired image. Reflection causes white spots in the image which affects the matching score due to missing / wrong data at that spot. The specular reflection can be identified by using an adaptive thresholding. The input image is successively thresholded to identify the most obvious reflection points in the image using a high pixel intensity threshold. Some of the area around the reflection is not captured during this step. The threshold is lowered by a value of 10 to obtain new threshold and the number of pixels is counted again. If the number of pixels detected is slightly increased, then the area of specular reflection patch is slightly improved. The threshold is further reduced and the process is repeated until the amount of specular reflection is slightly improved. This process is repeated in order to ensure that all reflections are identified.

3 Experimental Results

Images of publicly available CASIA-Iris v4 lamp [1] and in house collected IITK iris database are used in the testing process. The CASIA Lamp iris database consist of 16,212 eye images from 411 subjects. This database is collected in one session but in varying lighting conditions. The IITK iris database is collected within IITK mostly from the students. It has 20,420 eye images from 1021 subjects. The images are collected in two sessions. The time gap between the two sessions is about one week. Each eye has 5 images per session. The images are collected using a iris camera in Near Infra-red illumination.

Quality Estimation Analysis : Quality of an iris image has been assigned into a class with the help of a support vector machine. Higher quality images have higher score. The support vector machine is trained using 750 iris images from the CASIA database and 250 images from IITK database.

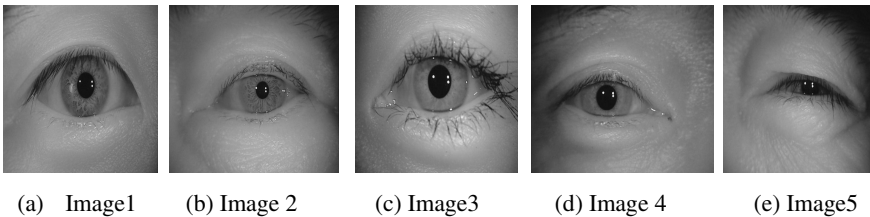
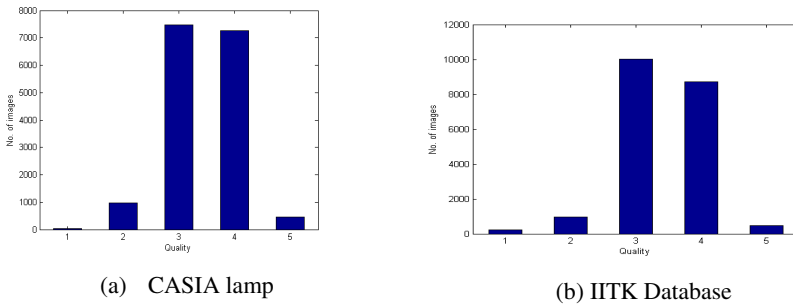


Fig. 4. Quality Estimation

Table 1. Estimated Quality Parameters and Quality Class for Images shown in Fig. 4

Image	Focus	Blur	Occlusion	Contrast	Dilation	Reflection	Quality
Image 1	0.2034	0.6783	0.8834	0.8873	0.8348	0.9865	5
Image 2	0.1985	0.6138	0.7289	0.8915	0.8517	0.9937	4
Image 3	0.1536	0.4974	0.7049	0.7760	0.7739	0.9103	3
Image 4	0.1648	0.5088	0.6790	0.7954	0.7856	1.0000	2
Image 5	0.1156	0.4067	0.2819	0.6659	0.7840	0.9589	1

In Fig. 4 some of the images from the database are shown. Their image quality parameters and the final quality score is shown in Table 1. It shows that as the quality of image degrades, its overall quality score also reduces. Image with higher occlusion has been given a poor score. The distribution of images among the quality classes is shown in Figure 5. Table 2 contains the number of images in each class for CASIA Lamp and IITK databases.

**Fig. 5.** Distribution of Image Quality Classes

Most of the images are found to be either in Class 3 or Class 4. The number of images in Class 1 (very poor quality) is very less as the quality of both the databases is good. The number of images in the highest class is also not very high in both databases.

Quality Based Performance Analysis: Methods of the quality estimation can be evaluated by comparing the performance of well established matching algorithm over different quality levels. A good quality estimation method improves the matching performance of the biometric system. The proposed method has been evaluated for its correctness by comparing the matching performance of Masek's [11] iris recognition system. The evaluation is being carried out by calculating the quality of all images in CASIA Lamp and IITK databases. CASIA Lamp database is split into two parts such that 8098 images are in the gallery set and 8113 images are in the query set. Similarly, IITK database is also divided into gallery and query sets with 10210 images each. All images in the gallery set are matched with all images in query set.

Quality of images tested is gradually raised using the pair-wise match quality which is defined as the minimum quality out of the two images considered in a matching pair as a reference metric.

Table 2. Quality Class Distribution

Image Class	No of Images	
	CASIA	IITK
1	27	223
2	971	966
3	7465	10032
4	7267	8727
5	461	472

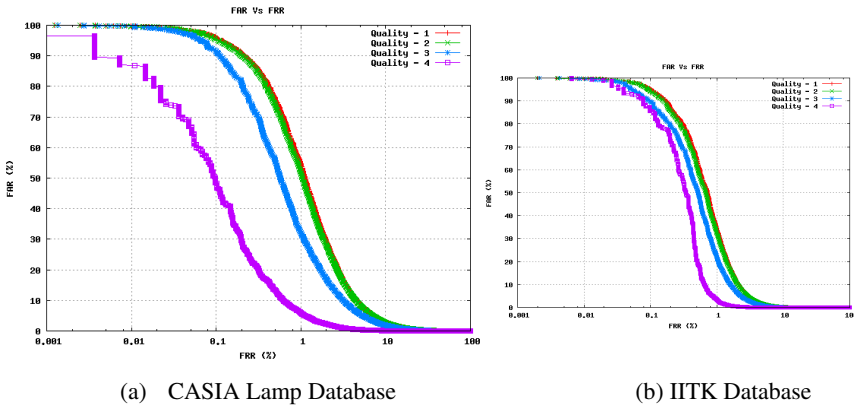


Fig. 6. Matching Performance against Quality

The graph of False Acceptance Rate (FAR) vs False Rejection Rate (FRR) is plotted in Figure 6 for each quality class. It can be seen that higher quality images in the database have higher matching performance. As the quality of the images is increasing, the False matches are reduced. The Equal Error Rate (EER) for CASIA lamp and IITK databases for different quality classes which are given in Table 3, are found to be reduced as the quality of the images is increased. The improvement in the EER is consistent with the increase in the quality of images.

Table 3. Equal Error Rate (EER) Values for Different Quality Classes

Image Class	Equal Error Rate (ERR)	
	CASIA	IITK
1	3.134	3.502
2	3.087	3.441
3	2.597	2.946
4	1.550	2.142
5	0.172	1.329

4 Conclusions

In this paper, an iris quality estimation method has been proposed that makes use of attributes like the focus of image, motion blur in image, occlusion, contrast, dilation and specular reflections. The estimation has been carried out by assigning them into quality based classes. It has been tested on the CASIA Lamp and IITK iris databases. It has been found that it has successfully obtained the quality of all images and the matching performance has also been improved with the increase in quality.

Acknowledgements. Authors like to acknowledge the support provided by the Department of Information Technology, Government of India to carry out this research work.

References

1. The casia lamp iris database, <http://iris.idealtest.org/>
2. Badrinath, G.S., Nigam, A., Gupta, P.: An Efficient Finger-Knuckle-Print Based Recognition System Fusing Sift And Surf Matching Scores. In: International Conference on Information and Communications Security, pp. 374–387 (2011)
3. Badrinath, G.S., Tiwari, K., Gupta, P.: An Efficient Palmprint Based Recognition System Using 1d-Dct Features. In: Huang, D.-S., Jiang, C., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2012. LNCS, vol. 7389, pp. 594–601. Springer, Heidelberg (2012)
4. Bendale, A., Nigam, A., Prakash, S., Gupta, P.: Iris segmentation using improved hough transform. In: Huang, D.-S., Gupta, P., Zhang, X., Premaratne, P. (eds.) ICIC 2012. CCIS, vol. 304, pp. 408–415. Springer, Heidelberg (2012)
5. Chen, Y., Dass, S.C., Jain, A.K.: Localized Iris Image Quality Using 2-D Wavelets. In: Proceedings of International Conference on Biometrics, pp. 373–381 (2006)
6. Daugman, J.: High Con_Dence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1148–1161 (1993)
7. Daugman, J.: How Iris Recognition Works. In: Proceedings of International Conference on Image Processing, pp. 33–36 (2002)
8. Ferzli, R., Karam, L.J.: A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (Jnb). IEEE Transactions on Image Processing 18(4), 717–728 (2009)
9. Kalka, N.D., Jinyu, Z., Schmid, N.A., Cukic, B.: Image Quality Assessment for Iris Biometric. In: Proceedings of SPIE. The International Society for Optical Engineering. Society of Photo-Optical Instrumentation Engineers (2006)
10. Kumar, J., Nigam, A., Prakash, S., Gupta, P.: An Efficient Pose Invariant Face Recognition System. In: International Conference on Soft Computing or Problem Soling, SocProS (2), pp. 145–152 (2011)
11. Masek, L.: Recognition of Human Iris Patterns for Biometric Identi_Cation. Technical report, University of Western Australia (2003)
12. Nigam, A., Gupta, P.: Finger Knuckleprint Based Recognition System Using Feature Tracking. In: Sun, Z., Lai, J., Chen, X., Tan, T. (eds.) CCBP 2011. LNCS, vol. 7098, pp. 125–132. Springer, Heidelberg (2011)
13. Nigam, A., Gupta, P.: A New Distance Measure for Face Recognition System. In: International Conference on Image and Graphics, ICIG, pp. 696–701 (2009)

14. Nigam, A., Gupta, P.: Comparing Human Faces Using Edge Weighted Dissimilarity Measure. In: International Conference on Control, Automation, Robotics and Vision, ICARCV, pp. 1831–1836 (2010)
15. Nigam, A., Gupta, P.: Iris Recognition Using Consistent Corner Optical Flow. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 358–369. Springer, Heidelberg (2013)
16. Shafer, G.: A Mathematical Theory of Evidence (1976)
17. Singh, N., Nigam, A., Gupta, P., Gupta, P.: Four Slap Fingerprint Segmentation. In: Huang, D.-S., Ma, J., Jo, K.-H., Gromiha, M.M. (eds.) ICIC 2012. LNCS, vol. 7390, pp. 664–671. Springer, Heidelberg (2012)
18. Tiwari, K., Arya, D.K., Gupta, P.: Palmprint Based Recognition System Using Local Structure Tensor and Force Field Transformation. In: Huang, D.-S., Gan, Y., Gupta, P., Gromiha, M.M. (eds.) ICIC 2011. LNCS, vol. 6839, pp. 602–607. Springer, Heidelberg (2012)
19. Wan, J., He, X., Shi, P.: An Iris Image Quality Assessment Method Based on Laplacian Of Gaussian Operation. In: Proceedings of Machine Vision Application (MVA) Conference. Citeseer, Tokyo (2007)
20. Wei, Z., Tan, T., Sun, Z., Cui, J.: Robust and Fast Assessment of Iris Image Quality. In: Proceedings of International Conference on Biometrics, pp. 464–471 (2006)

Age-Invariant Face Recognition Using Shape Transformation

Shubham Jain, Aditya Nigam, and Phalguni Gupta

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur
Kanpur 208016, India
{shubji,naditya,pg}@cse.iitk.ac.in

Abstract. This paper proposes a novel facial image transformation that minimizes the variation due to aging in facial features. It transforms a face image of an individual according to the probe image to register the facial features. The images are globally registered using Delaunay triangulation which reduces the linear variation in individual's appearance due to aging. Finally, weighted Local Binary Patterns are used to calculate the similarity between the registered images. The proposed approach has been tested on the publicly available FG-NET database [2] and a self-created BROWNS database. It is found to be robust to aging variation to a good extent. Gain in Rank-1 accuracies of 18.36 % and 17.37 % have been achieved on FG-NET and BROWNS database respectively.

Keywords: Biometrics, Face Recognition, Age-Invariant, AAM, LBP.

1 Introduction

Human face provides important information about an individual's identity. Easy availability of facial images and less cooperation of subject for capturing the same make face a well accepted biometric trait for person's recognition. However, the facial images suffer from variation due to pose, illumination, occlusion and aging because they degrade the facial features that can be extracted from the images. Variation due to pose, occlusion and illumination can be reduced to a certain extent by controlling the condition of image acquisition. But it is not the case with aging which is natural process. The process of aging differs for each individual and depends on different factors like gender, ethnicity, habits, lifestyle etc. [15]. Aging can reduce the usability of a facial database over time and render it useless for the purpose of identifying individuals. Therefore, compensating for aging variation in any facial database plays a key role in increasing the longevity of a face recognition system.

There exist several age simulation and modeling techniques which model the facial growth over a period of time to minimize the difference between probe and gallery images. Lanitis et al. [8] have proposed statistical models for face simulation. Training images have been used to learn the relationship between coded face representation and actual age of subjects. This relationship has been considered to estimate the age

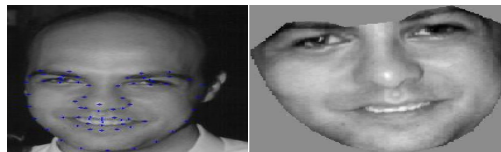
of an individual and to reconstruct the face at any age. Wang et al.[15] have obtained the texture and shape information of a face image using PCA and have used this information in reconstructing shape and texture at any particular age. Wu et.al. [15] have come up with a dynamic model to simulate wrinkles in 3D facial animation and skin aging. Skin deformations have been represented as plastic-visco-elastic processes and permanent wrinkles have been generated through a simulation of inelastic skin deformations. Geng et al.[4] learn a subspace of aging pattern based on the assumption that similar faces age in similar ways. The face representation is composed of face texture and the 2D shape represented by the coordinates of the feature points as in the Active Appearance Models [3]. Jain et al.[11] have proposed a learning approach for aging pattern based on PCA coefficients in separated 3D shape and texture, given a 2D database.

This paper presents a shape transformation approach to reduce the variation in facial features due to aging between two images of the same individual. The transformation is performed by extracting shape information from the face images by using a set of landmark points. We use these landmark points to find a Delaunay triangulation for transforming the gallery image according to the probe image. Finally, feature vectors are extracted using Local Binary Patterns to match the probe image with the transformed gallery image. The rest of the paper is organized as follows: Section 2 discusses the background required for the proposed age-invariance approach. Section 3 proposes a transformation method using Delaunay triangulation. The experimental results are analyzed in Section 4. Conclusions are given in the last section.

2 Background

2.1 Active Appearance Model(AAM) Search

Cootes et al. [3] have proposed a statistical approach to model facial appearance known as Active Appearance Models (AAM) which have been successfully used in various areas to model the shape and appearance of faces such as face recognition [13], facial expression recognition [13], age estimation and progression [8] and various other fields. The central problem of creating the combined model for an unseen image without any landmark points needs to be addressed. For this, a scheme of adjusting the model parameters is used so that the synthetically regenerated image from these parameters matches the original image as closely as possible.



(a) Landmarked Face (b) Shape Free Transform

Fig. 1. Landmarked Face and Shape Free Texture (Image from FG-NET Database [2])

This can be interpreted as an optimization problem in which the distance between a new image and the synthesized one obtained with the help of model parameters is minimized. If I is a column vector representing the image of size $n \times n$ and J is the corresponding vector which is obtained through the current model parameters then the difference vector can be defined as :

$$\delta_I = I - J \quad (1)$$

The magnitude of difference vector, $\Delta = \|\delta\|^2$ has to be minimized to locate the best match between the model and the image derived through the model parameters. As the number of model parameters is large, this is a high-dimensional optimization problem which is difficult to solve. But the solution to each of these is similar and this can be learned using the known model parameters for the images in the database. By using the a-priori knowledge of how to adjust the model parameters, an efficient algorithm can be used to search for new model parameters. The linear model can be used for defining the relationship between δ_I (the difference vector) and δ_c (the error in model parameters) as:

$$\delta_c = A \cdot \delta_I \quad (2)$$

where A is a vector of constants which can be obtained by the method of multivariate linear regression on a sample of known model displacements, δ_c , and the corresponding difference images, δ_I .

2.2 Delunay Triangulation [9]

Delaunay triangulation for a set of points P in a plane is a triangulation $T(P)$ such that every circum circle of a triangle $t(t \in T)$ is an *empty circle* which does not contain any other point $p(p \in P)$ inside itself. Thin triangles have large circum circles as compared to their area, Delaunay triangulation tends to avoid such triangles, i.e. maximize the minimum angle of each triangle in the triangulation. Given a set of control points or feature points, this property allows Delaunay triangulation to be used in modeling a terrain and further for applications like registering images.

3 Proposed Approach

This paper uses a transformation approach for the purpose of age-invariant face recognition. Instead of learning personalized aging patterns, an attempt has been made to transform each gallery image according to the shape of probe image. An individual's face undergoes a substantial change in shape during the process of aging. Therefore, reduction of these shape variation can be used to reduce the differences in appearance of various images of same individual. Since there exists shape variation between any two different individuals, this shape transformation tends to bring closer the images for same individual and separate further apart the images of different individuals. This assumption helps to design a better identification process. Figure 3 describes the transformation process for a sample probe and gallery image of different individuals. The proposed approach is divided into three steps: image transformation, feature extraction and matching.

3.1 Image Transformation

All gallery images are transformed according to the probe image. The proposed transformation consists of three steps: selection of Landmark points from an unseen probe image, Delaunay triangulation for gallery image, Transformation of gallery image. Landmark points are (or control points) marked on the image which has to be transformed as well as on the target image. These points are marked in a particular order on each image as shown in Figure 2(a), (b). This is done manually for all images in the database. For a new image which has to be matched against the database, these points can be found using AAM search as discussed in Section 2.1. The marked points are triangulated using Delaunay triangulation. The set of fold-over triangles are eliminated because such triangles do not map correctly to the target triangulation and distort the face image. After removing such triangles, Delaunay triangulation for all selected points for gallery image is determined as shown in Figure 2(c), (d). This triangulation is used to transform the gallery images to the shape of probe images. This is done by inferring an affine mapping from gallery image to probe image coordinates using the three vertices of each triangle [6]. This mapping is applied to the gallery images to *register* the shapes of the two images together as shown in Figure 2(f). These registered images are used for feature extraction and matching.

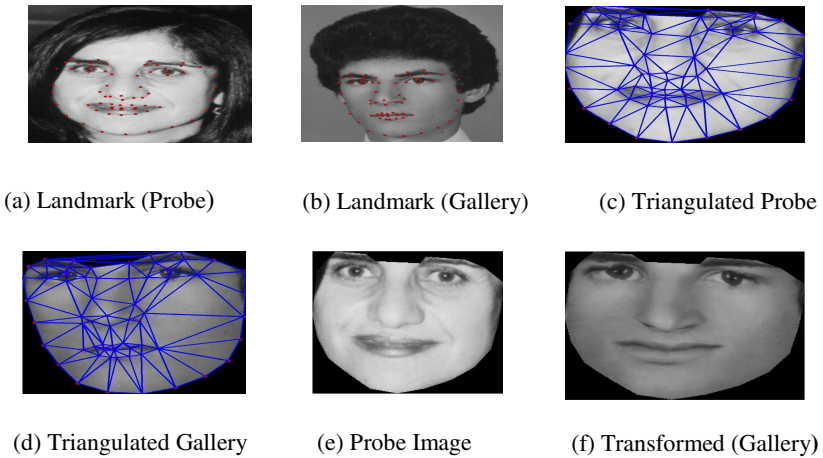


Fig. 2. Steps of proposed approach: Step 1:((a),(b)) Finding landmark points on probe and gallery images, Step 2:((c),(d)) Delaunay triangulation for probe and gallery image, Step 3:((e),(f)) Transforming gallery image according to shape of probe

3.2 Feature Extraction

The Local Binary Pattern (LBP) based features are calculated from the enhanced transformed gallery as well as probe images in order to compare them. Feature extraction consists of the following 3 steps: Image Enhancement, Regions of Interest (ROI) extraction and Local Binary Patterns (LBP) histogram calculation.

Image Enhancement: Transformed images are enhanced to make the features more prominent and thus it increases the discriminatory power of feature vector. In the proposed approach, images are enhanced using a two step process:

1. Blockwise Mean Subtraction: The transformed image is divided into 8×8 pixel size blocks and the average gray value of each block is subtracted from itself. This removes non-uniform illumination from the image.

$$I_i^{en} = I_i - \bar{I}_i \quad (3)$$

where I_i^{en} is the i^{th} block of enhanced image and I_i is the i^{th} block of original image. Figure 3(a) shows the transformed gallery image and Figure 3(b) shows the same image after mean subtraction.

2. Contrast Limited Adaptive Histogram Equalization (CLAHE): It enhances the facial features without increasing the background noise [16] as shown in Figure 3(c).

ROI Extraction. The facial image contains more useful information in some regions like eyes, nose, lips etc. as compared to other regions [8, 5]. Therefore for calculating feature vector only these regions are used. The location of these regions is calculated using the location information of landmark points for each image.

$$(x_{\min}^i, y_{\min}^i) = (\min(s_x^i) - b, \min(s_y^i) - b) \quad (4)$$

$$(x_{\max}^i, y_{\max}^i) = (\max(s_x^i) + b, \max(s_y^i) + b) \quad (5)$$

Where (x_{\min}^i, y_{\min}^i) and (x_{\max}^i, y_{\max}^i) are the bounding corners of the ROI for region i , $s^i \in S^i$ { S^i is the set of landmark points in region R_i }, $i \in \{1,2,3,4\}$, b is the padding (in number of pixels) around the landmark points. Its value is experimentally determined.

LBP Histogram Calculation. LBP histogram is calculated for both, transformed gallery image and probe image. For this purpose $LBP_{(8,1)}^{u2}$ operator is used as described in [1]. All the non-uniform patterns (patterns which have more than two bit transitions) are added to a single bin of the histogram. All four regions (left eye, right eye, nose and lips) are divided into $8 \times 8 = 64$ blocks. For each block, an LBP histogram is calculated using the decimal values of local binary patterns as labels. The concatenation of the histograms for each block in each region is considered as the feature vector of the image.

3.3 Feature Matching

The LBP histogram are compared using Chi square (χ^2) dissimilarity measure defined as :

$$\chi^2(H_{probe}^{R,B}, H_{gallery}^{R,B}) = \sum_{i=0}^{n-1} \frac{(H_{probe}^i - H_{gallery}^i)^2}{H_{probe}^i + H_{gallery}^i} \quad (6)$$

where n is the number of histogram labels, $n = P^2 - P + 3$ which is 1 + number of uniform patterns for P sampling points ($P = 8$ in this case), H_{probe}^i and $H_{gallery}^i$ are the values for i^{th} labels of probe and gallery histograms, respectively, $H_{probe}^{(R,B)}$ and $H_{gallery}^{(R,B)}$ are the histograms for block B of a region R for probe and gallery image, respectively. Using Eq. 6, the overall dissimilarity score for probe and gallery images can be expressed as

$$D(I_{probe}, I_{gallery}) = \sum_{R=1}^4 \sum_{B=1}^{64} \chi^2(H_{probe}^{R,B}, H_{gallery}^{R,B}) \quad (7)$$

where B refers to each block in a region R of the image. The distance measure D quantifies the feature level difference between the images. Lower value of D signifies that the images are similar and vice versa.

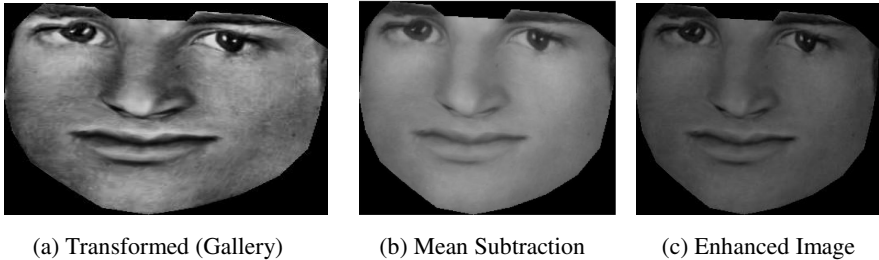


Fig. 3. Enhancement

4 Experimental Results

We have used the publicly available FG-NET [2] aging database for our experiments. FG-NET database contains scanned 1002 images of 82 subjects, varying in pose, expression and illumination. Each face image has 68 landmarked points as shown in Fig. 1. Another facial aging database is created from a book by Nixon and Galassi[10]. This book contains pictures of four sisters over a period of 36 years. These images have been scanned and face areas are cropped to obtain age-progressive images of four individuals with 36 images each. This database is henceforth referred to as the “BROWNS” database. The BROWNS database has much lesser pose variation and better quality images as compared to FG-NET database.

For the purpose of evaluating the proposed approach we have used a leave-one-out approach on FG-NET and BROWNS databases. One image from the database is matched against all other images and the performance is evaluated in terms of

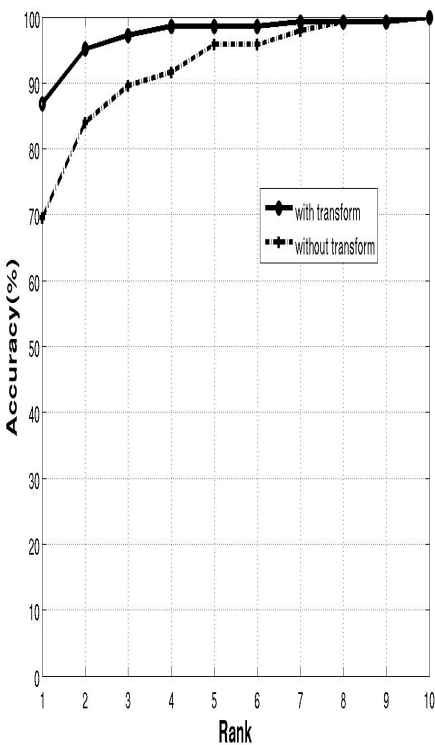
Rank-Identification Accuracy. This measure gives the probability of a genuine match being in the top R matches, where R is called the rank of identification accuracy.

$$Acc_{Identification}^R = \frac{\sum_{i=1}^N A(ident(i) \in top_match^R(i))}{N} \tag{8}$$

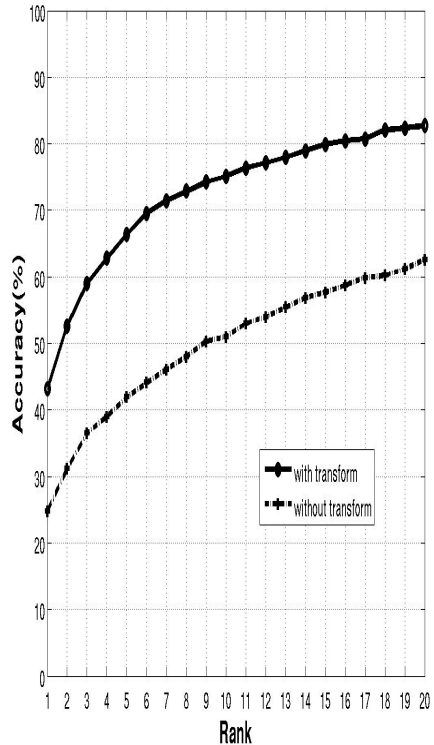
where N is the total number of images in database, $indent(i)$ is the individual identity of image i , $top_match^R(i)$ is the set of identities of the top R matching images with i^{th} probe image, and

$$A\{T\} = \begin{cases} 1 & \text{If } T \text{ is true} \\ 0 & \text{If } T \text{ is false} \end{cases}$$

The CMC curves for the FG-NET and BROWNS databases are shown in Figure 4(a),(b). A gain of 18.36 and 17.37 has been obtained in rank-1 accuracies for FG-NET and BROWNS database, respectively. Table 1 shows the comparison of the proposed approach with some other published results. The approach is found to give comparable performance with the existing approaches for FG-NET database.



(a) FG-NET database



(b) BROWNS database

Fig. 4. Rank Identification curves for the proposed approach

Table 1. Comparison of proposed approach with existing approaches

	Matching Algorithm	Database	Rank -1 Identification Accuracy	
			Original image	After aging model
Geng et al. (2007) [4]	Mahalanobis distance, PCA	FG-NET*	14.4	38.1
Jain et al. [11]	Face VACS	FG-NET**	26.4	37.4
		BROWNS	15.6	28.1
Proposed Approach	weighted LBP (leave-one-out)	FG-NET**	24.75	43.11
		BROWNS	69.44	86.81

* Used only a very small subset of the FG-NET database that contains 82 subjects

** Used all the subjects of FG-NET

5 Conclusions

In this paper, a novel approach has been presented for the problem of automatic age variance. Humans undergo significant changes in shape and appearance of face during process of aging. To reduce the effect of these variations on performance of face recognition systems, a shape in query image has been used. Feature vectors are extracted using weighted LBP. Extensive experiments on the standard FG-NET and BROWNS database allow us to conclude that the proposed approach gives comparable results with the other well known approaches. Further, through experiments we can also conclude that a system provides age-invariance to a certain extent, beyond which it fails to match images effectively.

Acknowledgements. Authors like to acknowledge the support provided by the Department of Information Technology, Government of India to carry out this research work.

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. Cootes, T.: The Fg-Net Aging Database, <http://sting.cycollege.ac.cy/~alanitis/fgnetaging/index.htm>
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
4. Geng, X., Zhou, Z.-H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(12), 2234–2240 (2007)

5. Gong, S., McKenna, S.J.: *Dynamic Vision: from Images to Face Recognition*, 1st edn. Imperial College Press, London (2000)
6. Goshtasby, A.: Piecewise Linear Mapping Functions for Image Registration. *Pattern Recognition* 19(6), 459–466 (1986)
7. Kang, H., Cootes, T.F., Taylor, C.J.: A Comparison of Face Verification Algorithms Using Appearance Models. In: *Proceedings of the 13th British Machine Vision Conference (BMVC 2002)*, vol. 2, pp. 477–486 (2002)
8. Lanitis, A., Taylor, C., Cootes, T.: Toward Automatic Simulation of Aging Effects on Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(4), 442–455 (2002)
9. Leach, G.: Improving Worst-Case Optimal Delaunay Triangulation Algorithms. In: *Proceedings of the 4th Canadian Conference on Computational Geometry*, vol. 15 (1992)
10. Nixon, N., Galassi, P.: *The Brown Sisters: Thirty-Three Years*. Museum of Modern Art (2007)
11. Park, U., Tong, Y., Jain, A.K.: Age-invariant Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(5), 947–954 (2010)
12. Scherbaum, K.: Prediction of Individual Non-Linear Aging Trajectories of Faces. *Computer Graphics Forum*. 26(3), 285–294 (2007)
13. Tang, F., Deng, B.: Facial Expression Recognition Using AAM and Local Facial Features. In: *Proceedings of the Third International Conference on Natural Computation*, vol. 2, pp. 632–635 (2007)
14. Wang, J., Shang, Y., Su, G., Lin, X.: Age simulation for Face Recognition. In: *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, vol. 3, pp. 913–916 (2006)
15. Wu, Y., Thalmann, N.M., Thalmann, D.: A Dynamic Wrinkle Model in Facial Animation and Skin Aging. *The Journal of Visualization and Computer Animation* 6(4), 195–202 (1995)
16. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization. In: *Graphics Gems iv*, pp. 474–485 (1994)

Low-Resolution Image Restoration Using the Combination Method of Sparse Representation and PDE Model

Li Shang¹ and Zhan-li Sun²

¹ Department of Communication Technology,
College of Electronic Information Engineering, Suzhou Vocational University,
Suzhou 215104, Jiangsu, China

² Department of Automation, College of Electrical Engineering and Automation,
Anhui University, Hefei 230601, Anhui, China
s10930@jssvc.edu.cn, zhlsun2006@yahoo.com.cn

Abstract. The stable solutions of traditional partial differential equations (PDE) can cause obvious step effects when PDEs are utilized to restore low-resolution images, and the quality of images restored is hardly worse. To solve this problem above-mentioned, a new low-resolution image restoration method, based on the combination method of sparse representation and PDE model based on an enhanced total variation model (ETVM), is proposed in this paper. The dictionary of sparse representation of images is learned by using the K-means based singular value decomposition (K-SVD) algorithm. For images with large noise variance or low-resolution, K-SVD has better denoising robustness. The guiding ideology of low-resolution image restoration is that the K-SVD algorithm is used first to reduce unknown noise existed in low-resolution images, and then the PDE model based on total variation (TV) are utilized to restore the results denoised obtained by K-SVD. In test, a human-made and a real low-resolution image, called millimeter wave (MMW) image, are respectively used to testify our method proposed. Further, compared it with algorithms of K-SVD and PDE, at the same time, the pick signal noise ratio (PSNR) criterion is used to measure restored human-made low-resolution images. Considering different noise variance for a human-made low-resolution image, and in terms of PSNR values and the vision effect of restored images, simulation results show that our method proposed here can efficiently restore low-resolution images, and behave certain theory meaning and practicality.

Keywords: Sparse representation, K- Singular Value Decomposition (K-SVD), Partial differential equation (PDE), Millimeter wave (MMW) image, Image restoration, Image denoising.

1 Introduction

The technology of Low-resolution image restoration has always been a challenge a question for study or discussion in image processing field to this day. It is well known

that some classical image restoration algorithms have been developed efficiently, such as inverse filtration [1], Wiener filtration [2], mid-value filtration and Gaussian filtration [3-4], the least square method without constraint or that with constraint [5-6], singular value decomposition (SVD) [7], wavelet transform [8], etc. These methods are usually divided into two types: the certain analysis and calculation method, and the statistical method. However, when using these methods to restore low-resolution images, some assumed conditions, for example, low-resolution images not only stratify generalized stationary conditions but also positive constrained conditions, must be in advance considered [3-4]. So, these method above-mentioned are hardly limited in application. Otherwise, the basis task of image restoration is to delete noise generated in image degrading, and meanwhile to retain details of images. But, it is hardly difficult in restraining noise and retaining detail simultaneously. Just to solve this problem, in this paper, sparse representation and partial differential equations (PDE) based on an enhanced total variation model (ETVM) are combined to restore low-resolution images [9-12]. As we all know, the theory of PDE is very simple and is easy to compute, at the same time, its convergence rate is very quick. But, when noise is much large, the PDE model's stable solutions can cause obvious step effects and the quality of image restoration is very worse. For this reason, sparse representation method [7, 13-15] is considered as the preprocessing of PDE in this paper. Sparse representation is an efficient signal (or image) representation method based on dictionary, and here the dictionary is learned by using K-means based singular value decomposition (K-SVD) algorithm [7]. Using atoms of the dictionary learned, a low-resolution image can be denoised efficiently. Utilizing the combination of K-SVD and PDE, a better restoration result of low-resolution may be obtained. Further, for human-made and real low-resolution images, respectively compared with K-SVD and PDE and using the pick signal noise ratio (PSNR) criterion to measure the quality of restored images, simulation results show that our method proposed (denoted by K-SVE+PDEs) is the best and behave certain research meaning and practicality.

2 Sparse Representation Method

2.1 Sparse Representation of Images

Sparse representations are representations that contain most or all information of a signal with a linear combination of a small number of elementary signals called atoms. Often, these atoms are chosen from a so called over-complete dictionary. Let the matrix $D \in \mathfrak{R}^{N \times K}$ be an over-complete dictionary of K prototype atoms, and supposed a vector $x \in \mathfrak{R}^N$ can be represented as a sparse linear combination of these atoms. Thus, this signal x can be approximately written as $x \approx Ds$, satisfying $\|x - Ds\|_p \leq \varepsilon$, where $s \in \mathfrak{R}^K$ is a vector with very few ($\ll K$) nonzero entries [8]. In approximation methods, typical norms used for measuring the deviation are the l^p norms ($p = 1, 2, \dots, \infty$). Generally, It is concentrated on the case of $p = 2$. If $N < K$ and D is a full-rank matrix, an infinite number of solutions are available for the

problem, hence constraints on the solution must be set. The solution with the fewest number of nonzero coefficients is certainly an appealing representation. This sparsest representation is the solution of either [13]

$$(p_0) \min_s \|s\|_0 \text{ subject to } x = Ds \text{ .} \tag{1}$$

or

$$(p_{0,\epsilon}) \min_s \|s\|_0 \text{ subject to } \|x - \tilde{x}\| = \|x - Ds\|_2 \leq \epsilon \text{ .} \tag{2}$$

where the symbol $\|\cdot\|_0$ is the l^0 norm, counting the nonzero entries of a vector and \tilde{x} is the approximate value of x .

For an image, the idea of sparse representation is the same as that of a signal. The goal of an image’s sparse representation is in fact to find a dictionary D , and then this image can be expressed by utilizing a few non-zero elements $\{d_{ii}\}$. Here, Let X denote the matrix of an image, and S denote the feature coefficient matrix of this image. And then, the optimized object function of sparse representation for an image can be written as follows

$$\min_s \|S\|_0 \text{ s.t. } \left\| X - \sum_k^K s_k d_k \right\|_2 \leq \epsilon \text{ .} \tag{3}$$

where the parameter K is the number of atoms, d_k is the k th atom of the dictionary D , and s_k is the k th column vector of coefficient matrix S .

2.2 Dictionary Learning of Sparse Representation

The usual dictionary learning algorithm of sparse representation is K-SVD method. K-SVD is flexible and works in conjunction with any pursuit algorithm, and it is designed to be a truly direct generalization of the K-Means. Assumed that the objective function is the following form:

$$\min_{D,S} J(D,S) = \|X - DS\|_2^2 \text{ subject to } \forall_i : \|s_i\|_0 \leq T_0 \text{ .} \tag{4}$$

where T_0 is a positive threshold value. First, considering the sparse coding stage, the penalty term can be rewritten as

$$\min_{D,S} J(D,S) = \sum_{i=1}^N \|x_i - Ds_i\|_2^2 \text{ .} \tag{5}$$

Therefore the problem posed in Equation (5) can be decoupled to N distinct problems of the form:

$$\min_{s_i} \left\{ \|x_i - Ds_i\|_2^2 \right\} \text{ subject to } \|x_i\|_0 \leq T_0, i = 1, 2, 3, \dots, N \text{ .} \tag{6}$$

This problem is adequately addressed by the pursuit algorithms, and if T_0 is small enough, the solution of Equation (5) is a good approximation to the ideal one that is numerically infeasible to compute. And consider the error between the original image data and its approximate values, Equation (5) can be written as

$$\|X-DS\|_2^2 = \left\| X - \sum_{j=1}^K s_j d_j \right\|_2^2 = \left\| \left(X - \sum_{j \neq k} s_j d_j \right) - s_k d_k \right\|_2^2 = \|E_k - s_k d_k\|_2^2. \tag{7}$$

where the matrix E_k stands for the error for all the N examples when the k th atom is removed. However, because the vector s_k is very likely to be filled, in such an update of d_k , we don't enforce the Sparsity constraint. To solve this problem, we define ω_i as the group of indices pointing to examples $\{x_i\}$ that use the atom d_k , thus, $\omega_k = \{i | 1 \leq i \leq K, s_k \neq 0\}$. Further, we define B_k as a matrix of size $N \times |\omega_k|$, with ones on the $(\omega_k(i), i)$ -th entries, and zeros elsewhere. Define $\tilde{s}_k = B_k s_k$, $\tilde{X}_k = X B_k$, $\tilde{E}_k = E_k B_k$, and parameters of \tilde{s}_k , \tilde{X}_k and \tilde{E}_k are respectively the shrinkage results of s_k , X and \tilde{E}_k removed zero entries. \tilde{X}_k creates a matrix of size $n \times |\omega_k|$ that includes a subset of the examples that are currently using the d_k atom. \tilde{E}_k has the same size of \tilde{X}_k , and it is a selection of error columns that correspond to examples that use the d_k atom. With this notation, the solution of Equation (7) \tilde{x}_k has the same support as the original x_k . This is equivalent to the minimization of the following formula

$$\|E_k B_k - d_k (s_k)^T B_k\|_2^2 = \|\tilde{E}_k - d_k (\tilde{s}_k)^T\|_2^2. \tag{8}$$

Then, Equation (8) is done directly via SVD. Taking the restricted matrix \tilde{E}_k , SVD decomposes it to $\tilde{E}_k = U \Delta V^T$. Here, the vector \tilde{d}_k is defined as the first column of U , and \tilde{s}_k is defined as the first column of V multiplied by $\Delta(1,1)$. This algorithm is called as K-SVD to parallel the name K-Means. While K-Means applies K computations of means to update the codebook, the K-SVD obtains the updated dictionary by K-SVD computations, each determining one column.

3 The ETVM Based PDE Algorithm

Applications of the PDE models can be widely found in a broad range of image restoration tasks such as denoising, enhancement, inpainting and so on [18-19]. Currently, there have been developed many important nonlinear PDEs models. The model used widely is the variational PDE method [19]. In the variational or "energy" based models, nonlinear PDEs emerge as one derives their formal Euler-Lagrange equations, or tries to locate the local or global minima by the gradient descent method.

Assumed that u is the original true image, H is a blurring operator and n is noise signal. Then, a low solution image u_0 is described as the formula of $u_0|_{\Omega} = (Hu+n)|_{\Omega}$. The goal of inpainting is to reconstruct u as faithfully as possible form $u_0|_{\Omega}$. The data model is explicitly given by

$$E[u_0|u, \Omega] = \frac{1}{Q} \int_{\Omega} (Hu - u_0)^2 dx \quad (9)$$

Therefore, the quality of an inpainting model crucially depends on the prior model or the regularity energy $E[u]$. Considering the total variational (TV) prior model $E[u] = \int_{\Omega} |Qu|$ and the ordinary Sobolev norm, and in combination with the data model (9), the total variation minimization problem is considered:

$$\min_u \left(\alpha \int_{\Omega} |\nabla u| dx dy + \frac{1}{2} \|Hu - u_0\|_{L^2}^2 \right) \quad (10)$$

where $\alpha > 0$ is the penalty parameter, $\nabla u = \sqrt{u_x^2 + u_y^2}$, and Ω is the sustainable region of images. The corresponding Euler-Lagrange equation for Eqn. (4) is written as:

$$-\alpha \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) + H^*(Hu - u_0) = 0 \quad \text{in } \Omega \quad (11)$$

where ∇ is the Laplace operator, and H^* is the adjoint operator of H with respect to standard L_2 inner product. This idea gives a rigorous mathematical tool to introduce nonlinear diffusion filters in the image restoration. However, the basic TV model based PDE methods can show some drawbacks unless the governing equations are both incorporating appropriate parameters and discretized by suitable numerical schemes. Therefore, in this paper, the ETVM is used to restore images. The object function of ETVM is defined as follows:

$$\min_u \mathcal{F}_{\varepsilon,p}(u) \quad \mathcal{F}_{\varepsilon,p}(u) = \int_{\Omega} |\nabla^{\varepsilon} u|^p dx + \lambda \|u_0 - u\|^2 \quad (12)$$

where $\lambda \geq 0$ is a parameter (Lagrange multiplier) and $\varepsilon \geq 0$ is a regularization parameter:

$$|\nabla^{\varepsilon} u| = (u_x^2 + u_y^2 + \varepsilon^2)^{1/2} \quad (13)$$

Noted that when $p = 1$ and $\varepsilon = 0$, the first term in $\mathcal{F}_{0,1}(u)$ is namely the TV model. The Euler-Lagrange equation of Equation (13) is the following formula:

$$u_t = \nabla \cdot \left(\frac{\nabla u}{|\nabla^{\varepsilon} u|^{2-p}} \right) + \beta (u_0 - u) \quad (14)$$

where an artificial time t is introduced. However, Equation (14) often introduces a certain amount of blur during the process of noise removal. To avoid such a non-necessary diffusion, an anti-diffusion strategy, which selects $p = 1 - \omega$ ($\omega > 0$) and allows β to be variable, i.e., $\beta = \beta(s, t)$, is suggested. Thus, the ETVM is written as

$$u_t = \nabla \cdot \left(\frac{\nabla u}{|\nabla^\varepsilon u|^{1+\omega}} \right) + \beta(s, t)(u_0 - u) . \tag{15}$$

Thus, utilizing Equation (15) for the data set preprocessed by K-SVD algorithm, the restoration task of low-resolution images can be successfully implemented.

4 Experimental Results and Analysis

4.1 Denoising Idea of K-SVD

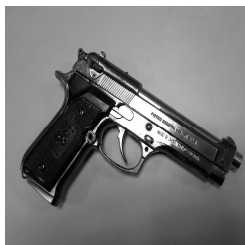
First, denoising results of low-resolution images of K-SVD method were discussed. In test, our method was testified by using a human-made and real low-resolution image respectively. The artificial low-resolution image generated by Barbara was contaminated by Gaussian noise with 20 noise variance and shown in Fig.1 (a). The real one shown in Fig.1 (b) is a millimeter wave (MMW) image with 41×41 pixels. The MMW image was generated by the State Key Lab. of Millimeter Waves of Southeast University, which is our cooperation group in the MMW image’s research.



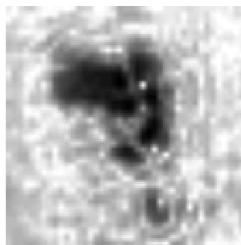
(a) Original Barbara image



(b) Degraded Barbar imge (20 noise level)



(c) Original toy gun image



(d) MMW image

Fig. 1. Original images and corresponding low-resolution images

In learning the dictionary D , the low-resolution image was sampled randomly sub-windows of 8×8 pixels 5000 times, and each sub-window was converted into one column, thus, a set with the size of 64×5000 was obtained. Using K-SVD to this set, the denoise process and dictionary learning process can be implemented simultaneously, and the optimized object function is written as

$$\arg \min_{S_{ij}, D, U} \left[\lambda \|F - \hat{F}\|_2^2 + \sum_{i,j} \mu_{ij} \|s_{ij}\|_0 + \sum_{i,j} \|D s_{ij} - R_{ij} F\|_2^2 \right]. \quad (16)$$

where \hat{D} and \hat{F} are respectively the estimation values of the dictionary D and original image patch matrix F , \hat{s}_{ij} denotes the coefficient vectors (located in the location $[i, j]$ of denoised image \hat{F}) of sparse representation on the dictionary \hat{D} obtained by K-SVD, μ_{ij} can be obtained in the learning of D , and $R_{ij}U$ denotes extractive coordinate $[i, j]$ from the original image U . For the degraded Barbara image, setting the redundancy rate to be 4 and the number of atoms is 64×256 , the dictionary D learned by K-SVD was shown in Fig.2 (a), and for comparison, the dictionary learned of discrete cosine transform (DCT) was also shown in Fig.2 (b).

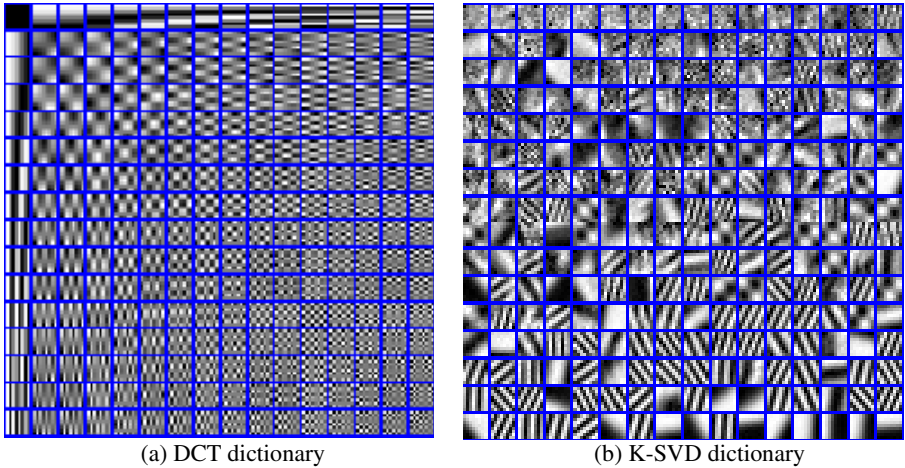


Fig. 2. Dictionary of degraded Barbara image with 20 noise level

4.2 Results of Image Restoration

First, the restoration result of the artificial low-resolution image was discussed to testify our method (here denoted by K-SVD+PDE, and PDE model was used the form of ETVM model). Considering different noise variance σ and the pink signal noise ratio (PSNR) measurement, PSNR values of denoised Barbara versions obtained by

K-SVD+PDE were listed in Table 1. And for comparison, those of denoised results, obtained by algorithms of K-SVD and PDE, were also shown in Table 1. At the same time, considering the paper’s length, only denoised images of degraded Barbara with 20 noise level, obtained by the three algorithms above-mentioned, were shown in Fig.3 (a) to Fig.3 (b).

Table 1. PSNR values of restored Barara images using different algorithms

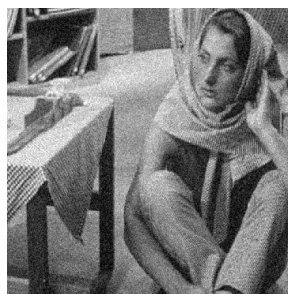
Algorithms Noise variance	KSVD+PDE	KSVD	PDE	Noise images
5	29.632	24.833	24.726	24.443
10	29.074	24.680	23.583	23.240
20	27.769	24.262	23.285	20.287
25	27.158	24.076	20.684	18.930
30	25.964	23.789	19.036	17.664
40	24.983	23.278	17.542	15.558
50	24.175	22.868	15.863	13.810
100	23.682	21.491	9.868	8.065
150	21.623	20.543	6.043	4.559



(a) K-SVD+PDE



(b) K-SVD



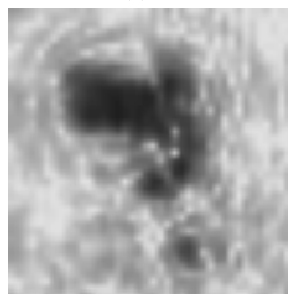
(c) PDE



(d) K-SVD+PDE



(e) K-SVD



(f) PDE

Fig. 3. Restored images Of different algorithms

According to experimental data in Table 1, it is clear to show that, to some extent, K-SVD and PDE respectively have distinct denoising effect, and K-SVD method is clearly better than PDE with the increasing of the noise level, moreover, our K-SVD+PDE method hardly outperforms each of the two methods. At the same time, it is noted that, despite of the algorithm used, the smaller the noise variance is, the larger the corresponding PSNR value is. These experimental results forcefully testify that our algorithm is highly effective in denoising low-resolution images, and can be further used to denoise a MMW image, i.e., the real low-resolution image. The restoration results of the MMW image obtained by different algorithms were shown in Fig.3 (d) to Fig.3 (e). Clearly, the denoising visual effect of the MMW image by using our K-SVD+PDE is the best in the discussed algorithms.

5 Conclusions

A novel low-resolution image restoration method using the combination algorithm of K-SVD and ETVM based PDE (i.e., K-SVD+PDE) is discussed in this paper. K-SVD is very efficient in image denoising, and here it is used to preprocess much and unknown noise existed in low-resolution images. In test, the restoration effect of our method proposed here is first testified by degraded Barbara image, which simulates an artificial low-resolution image. And the PSNR criterion is used to measure the quality of those images restored. Simultaneously, to show off our method, PSNR values of restoration results obtained by K-SVD and PDE are also given. After this, a real one is discussed by our method. Experimental results prove that the K-SVD+PDE method proposed by us is truly efficient and applied in low-resolution image restoration task.

Acknowledgement. This work was supported by the National Nature Science Foundation of China (Grant No. 60970058), the Nature Science Foundation of Jiangsu Province of China (No. BK2009131), the Innovative Team Foundation of Suzhou Vocational University (Grant No. 3100125), and the “333 Project” of Jiangsu Province.

References

1. Banham, M.R., Katsaggelos, A.K.: Digital image restoration. *IEEE Signal Proc. Magazine* 14, 24–41 (2009)
2. Hiroko, F., Shintaro, E., Testuya, S.: Image restoration via Wiener filtering in the frequency domain. In: *The 8th WSEAS International Conference on Signal Processing, Robotics and Automation*, Cambridge, England, UK, vol. 5(2), pp. 63–73 (2009)
3. Strang, G., Nguyen, T.: *Wavelet and Filter Banks*. Wellesley-Cambridge Press (1997)
4. Andrews, H.C., Hunt, B.R.: *Digital image restoraton*. Prentice Hall Press (1990)
5. Peng, X., Li, R., Yang, L.: Least square-based vehicle position estimation algorithm. *Journal on Communications* 31(8), 9–15 (2010)
6. Bojarczak, P., Osowski, S.: Denoising of image-A comparison of different filtering approaches. *WSEAS Trans. Computers* 3(3), 738–743 (2004)

7. Aharon, M., Elad, M.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54(11), 4311–4322 (2006)
8. Chang, S., Yu, B., Vetterli, M.: Adaptive wavelet threshold for image denoising and compression. *IEEE Transaction on Image Processing* 9(9), 1532–1546 (2000)
9. Chen, T., Shen, J.: Image processing and analysis: variational, PDE, wavelet, and stochastic methods. SIAM Publisher, Philadelphia (2005)
10. Bar, L., Sochen, N.A., Kiryati, N.: Image deblurring in the presence of salt-and-pepper noise. In: Kimmel, R., Sochen, N.A., Weickert, J. (eds.) *Scale-Space 2005*. LNCS, vol. 3459, pp. 107–118. Springer, Heidelberg (2005)
11. Liu, R., Lin, Z., Zhang, W., Su, Z.: Learning PDEs for image restoration via optimal control. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part I*. LNCS, vol. 6311, pp. 115–128. Springer, Heidelberg (2010)
12. Li, B., Su, Z.: An adaptive PDE image processing method based on L_p norm. *ACTA Automatica Sinica* 34(8), 849–853 (2008)
13. Chun, H.Z., Lei, Z.: Metasample-based sparse representation for tumor classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8(5), 1273–1282 (2011)
14. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, J.-D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.-L., Schumaker, L. (eds.) *Curves and Surfaces 2011*. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012)
15. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54(11), 4311–4322 (2006)

An Improved Image Corner Matching Approach

Bijin Yan, Fanhuai Shi, and Jiguang Yue

Dept. of Control Science & Engineering, Tongji University, China

Abstract. Accurate and reliable image corner matching plays an important role in various applications in computer vision, such as motion tracking, object recognition and 3D reconstruction. This paper proposes an improved image corner matching method based on both blob information and local neighbor similarity. The similarity measurement between two interest corner points is based on corresponding SIFT information, and can therefore help to achieve robust corner matching against rotation and scale changes. Experimental results demonstrate that the improved method is effective for image corner matching with significant rotation and scale changes as well as other common image corner matching conditions.

Keywords: corner matching, rotation and scale invariant, SIFT.

1 Introduction

Corner feature is one of the distinctive and easily detectable image features carrying good physical interpretations [1]. Image corner matching is an important task in computer vision and plays an essential role in various applications, such as motion tracking, object recognition and 3D reconstruction, so the accurate and reliable corner matching of two uncalibrated images differing significantly in rotation and scale is a meaningful task.

The representative traditional method for image corner matching is the normalized cross-correlation method [2]. The basic idea of traditional correlation-based approaches is to determine a candidate corner point match by comparing the similarity of the corresponding regions surrounding two corner points. The normalized cross-correlation method is simple, effective and easily hardware implemented, but unfortunately sensitive to rotation and scale changes. A newly modified normalized cross-correlation method is proposed in [3], which achieve robustness to rotation and scale changes. However, the interest corners are not exactly localized on the edge of the image.

Various kinds of features have been proposed for different applications in computer vision [4] [5]. Due to the robustness against rotation and scale changes, blob feature has gained wide application in many computer vision tasks. SIFT [6] has been proved to be reliable and highly-invariant against image rotation and scale changes.

A new hybrid approach has been proposed for robust corner matching [7]. Its main idea is to integrate SIFT, which can provide structure information, and therefore

achieve robustness against rotation and scale changes, with accurate localization of corner features. But there still exists disadvantage in the approach. Since only one SIFT feature is assigned to each corner feature, SIFT may not be able to provide sufficient structure information to determine a candidate corner match.

This paper presents an improved image corner matching approach based on both blob information based on SIFT and local neighbor similarity, which can automatically achieve robust corner matching between two uncalibrated images with significant rotation and scale changes. Our approach works as follows: First, interest corner points and SIFT feature points are detected for two images separately. Then, SIFT feature point correspondences between two images are established, and multiple SIFT feature points are assigned to each corner point. Finally, by computing the similarity measurement of two interest corner points based on blob information and local neighbor similarity, the corner point matches are established.

The remainder of the paper is organized as follows. Section 2 describes how to extract interest corner points and assign SIFT feature points to each corner point. Section 3 discusses the details of our proposed corner matching algorithm. Section 4 presents some experimental results on real images. Section 5 concludes the paper.

2 Interest Corner Point Detection and SIFT Feature Point Assignment

In this section, we will demonstrate how to use multiple SIFT feature points to provide sufficient structure information for each interest corner point.

Interest corner point detector used in our approach is enhanced Curvature Scale Space (CSS) corner point detector [8], which defines the local maxima of the absolute value of curvature as the corner points based on curvature function [9]. Other corner point detectors, such as Harris detector [10] and its modifications [11] [12], are also suitable for interest corner point detection in our approach.

SIFT feature is then chosen to provide reliable blob information in our method. First, SIFT feature points are extracted for both of the two images respectively. Then, initial SIFT feature point matches are established by comparing the vector angles of all possible matches. Finally, false SIFT feature matches are rejected based on the strategy that tends to exclude a match if the ratio of the nearest vector angle to the second nearest vector angle is above a specific threshold.

For each interest corner point, we assign a certain amount of SIFT feature points, for example, $N = 3$ SIFT feature points in our paper, which are closer to the interest corner point comparing with other SIFT feature points, to provide sufficient blob information around the interest corner point. By computing the normalized coordinates of each corner point with respect to corresponding SIFT feature points respectively, each corner point can be localized relative to corresponding SIFT feature points, and therefore can achieve rotation and scale invariance.

3 Corner Matching Based on Blob Information and Local Neighbor Similarity

Our approach defines the matching score based on blob information and local neighbor similarity to describe the similarity measurement between two interest corner points. The definition and calculation of the matching score based on blob information and local neighbor similarity are introduced in section 3.1 and 3.2 respectively. The details of matching process are discussed in section 3.3.

3.1 Matching Score Based on Blob Information

In our approach, multiple SIFT feature points are assigned to each corner points to provide sufficient blob information. First, we discuss how to compute the matching score based on blob information in the condition where only one SIFT feature point is assigned to each corner point, and then generalize to the common condition where N SIFT feature points are assigned to each corner point.

Let $c_1 = (x_1, y_1)$ be an interest corner point which a SIFT feature point $s_1 = (x_{s_1}, y_{s_1})$ is assigned to in one image and $c_2 = (x_2, y_2)$ be an interest corner point which a SIFT feature point $s_2 = (x_{s_2}, y_{s_2})$ is assigned to in the other image, σ_1, θ_1 and σ_2, θ_2 be the characteristic scale and dominant direction of the SIFT feature point s_1 and s_2 respectively, $c_1(x_1, y_1)$, $c_2(x_2, y_2)$, $s_1(x_{s_1}, y_{s_1})$ and $s_2(x_{s_2}, y_{s_2})$ are image coordinates. For the corner point c_1 in the first image, the relationship between c_1 and the local SIFT coordinate system centered at the corresponding SIFT feature point s_1 is showed in Figure 1.

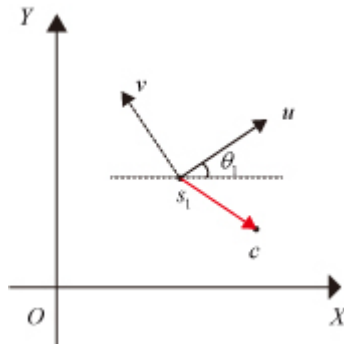


Fig. 1. Relationship between corner point c_1 and the local SIFT coordinate system centered at the corresponding SIFT feature point s_1

We can compute the normalized coordinates $\tilde{c}_1 = (\tilde{x}_1, \tilde{y}_1)$ in the SIFT coordinate system as follows:

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{y}_1 \end{bmatrix} = \frac{1}{\sigma_1} \begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{bmatrix}^T \begin{bmatrix} x_1 - x_{s1} \\ y_1 - y_{s1} \end{bmatrix} \quad (1)$$

Similarly for the corner point c_2 in the second image, we can also obtain the normalized coordinate $\tilde{c}_2 = (\tilde{x}_2, \tilde{y}_2)$ in the SIFT coordinate system centered at the corresponding SIFT feature point s_2 , which can be expressed as follows:

$$\begin{bmatrix} \tilde{x}_2 \\ \tilde{y}_2 \end{bmatrix} = \frac{1}{\sigma_2} \begin{bmatrix} \cos(\theta_2) & -\sin(\theta_2) \\ \sin(\theta_2) & \cos(\theta_2) \end{bmatrix}^T \begin{bmatrix} x_2 - x_{s2} \\ y_2 - y_{s2} \end{bmatrix} \quad (2)$$

Obviously we can use the cosine of the vector angle between \tilde{c}_1 and \tilde{c}_2 to determine the similarity between c_1 and c_2 :

$$\cos \langle \tilde{c}_1, \tilde{c}_2 \rangle = \frac{\tilde{x}_1 \tilde{x}_2 + \tilde{y}_1 \tilde{y}_2}{\sqrt{\tilde{x}_1^2 + \tilde{y}_1^2} \sqrt{\tilde{x}_2^2 + \tilde{y}_2^2}} \quad (3)$$

The matching score based on blob information can be defined as follows:

$$Score_B = \frac{\cos \langle \tilde{c}_1, \tilde{c}_2 \rangle + 1}{2} \quad (4)$$

which increases from 0 to 1 with the increase of similarity based on blob information between two interest corner points.

Consider the condition where the SIFT features $\{s_{11}, s_{12}, \dots, s_{1N}\}$ and $\{s_{21}, s_{22}, \dots, s_{2N}\}$ are assigned to interest corner points c_1 and c_2 respectively. Let $\tilde{c}_{1k} = (\tilde{x}_{1k}, \tilde{y}_{1k})$ be the normalized coordinates of interest corner points c_1 relative to the SIFT feature point s_1 and $\tilde{c}_{2k} = (\tilde{x}_{2k}, \tilde{y}_{2k})$ be the normalized coordinates of interest corner points c_2 relative to the SIFT feature point s_{2k} , where $k = 1, 2, \dots, N$. Similarly, we can build the matching score based on blob information as follows:

$$Score_B = \frac{\sum_{k=1}^N \lambda_k \cos \langle \tilde{c}_{1k}, \tilde{c}_{2k} \rangle + 1}{2} \quad (5)$$

which increases from 0 to 1 with the increase of similarity based on blob information between two interest corner points, where $\cos < \tilde{c}_{1k}, \tilde{c}_{2k} >$ is the cosine of the vector angle between \tilde{c}_1 and \tilde{c}_2 , the factor λ_k is defined as follows:

$$\lambda_k = \frac{1}{\|\tilde{c}_{1k}\| + \|\tilde{c}_{2k}\|} \sum_{i=1}^N \frac{1}{\|\tilde{c}_{1i}\| + \|\tilde{c}_{2i}\|} \tag{6}$$

where $\|\bullet\|$ is the norm of a vector.

3.2 Matching Score Based on Local Neighbor Similarity

In our paper, the correlation between two interest corner points is computed to compare neighbor similarity. To achieve the rotation and scale invariance, the orientation of the correlation window is determined based on the dominant directions of the corresponding SIFT features.

First we consider the condition where only one SIFT feature point is assigned to each corner point. Let's Assume $c_1 = (x_1, y_1)$, $c_2 = (x_2, y_2)$, $s_1 = (x_{s1}, y_{s1})$, $s_2 = (x_{s2}, y_{s2})$ and $\sigma_1, \theta_1, \sigma_2, \theta_2$ defined the same as in section 3.1. Without loss of generality, we assume $\sigma_1 > \sigma_2$, and define the scale ratio $r = \frac{\sigma_1}{\sigma_2}$ and the angle difference $\theta = \theta_1 - \theta_2$. W_1 and W_2 are two correlation windows of the size $(2w+1) \times (2w+1)$ centered on corner points c_1 and c_2 respectively, where w is a constant. W_1 is rotated by $|\theta|$ around c_1 . W_1 and W_2 can be represented by two arrays of pixel intensities A and B as follows:

$$\begin{aligned} A_{uv} &= I_1(x_1 + ru \cos \theta + rv \sin \theta, x_1 + rv \cos \theta + ru \cos \theta - rv \sin \theta) \\ B_{uv} &= I_1(x_2 + u, y_2 + v) \end{aligned} \tag{7}$$

Then the correlation between c_1 and c_2 can be expressed as follows:

$$Score_{NS} = \cos < A_{uv}, B_{uv} > \tag{8}$$

Then we consider the common condition where N SIFT feature points $\{s_{11}, s_{12}, \dots, s_{1N}\}$ and $\{s_{21}, s_{22}, \dots, s_{2N}\}$ are assigned to interest corner points c_1

and c_2 respectively. For $k = 1, 2, \dots, N$, two arrays of pixel intensities A_{1k} and B_{1k} can be computed Similarly. Therefore the matching score based on local neighbor similarity can be expressed as follows:

$$Score_{NS} = \sum_{k=1}^N \lambda_k \cos \langle A_{1k}, B_{2k} \rangle \quad (9)$$

where λ_k has been defined in section 3.1. It increases from 0 to 1 with the increase of the local neighbor similarity between two interest corner points.

3.3 Matching Process

The matching score between two an interest corner point pair is defined as the sum of the matching score based on blob information and local neighbor similarity. In order to extract as many initial corner point matches as possible, an iteration process strategy is applied in our method.

Suppose there are m interest corner points in the first image and n interest corner points in the second image. Consider the matching score matrix $S_1 \in M_{m,n}$, which is demonstrated in Figure 2, where its element s_{ij} stands for the matching score between the i -th interest corner point p_i in the first image and the j -th interest corner point q_j in the second image. In order to establish one to one interest corner point correspondence, two interest corner points will be accepted as a match only if their matching score s_{ij} is both the greatest element in its row and column. By employing this strategy, the set of interest corner point matches Ω_1 can be established. Then, elements in the corresponding rows and columns with respect to Ω_1 in the matrix S_1 will be deleted to form a new matching score matrix S_2 . For example, elements in the i -th row and the j -th column will be deleted if a corner point match $(p_i, q_j) \in \Omega_1$. Then the set of interest corner point matches Ω_2 can be build based on the matrix S_2 . Obviously the k -th iterative procedure can be expressed as follows:

1. build a set of corner point matches Ω_k based on the matching score matrix S_k ;
2. Update the matrix S_k to form a new matching score matrix S_{k+1} .

	q_1	• • •	q_j	• • •	q_n
p_1	s_{11}	• • •	$s_{1,j}$	• • •	$s_{1,n}$
•	•	•	•	•	•
•	•	•	•	•	•
p_i	$s_{i,1}$	• • •	$s_{i,j}$	• • •	$s_{i,n}$
•	•	•	•	•	•
•	•	•	•	•	•
p_m	$s_{m,1}$	• • •	$s_{m,j}$	• • •	$s_{m,n}$

Fig. 2. The matching score matrix S_1

After the iteration procedure, an initial set of interest corner point matches is established. But the initial set of interest corner point matches usually contains some false matches. In our paper, an initial corner match will be consider as a false match if either of its matching score based on blob information or local neighbor similarity is less than the certain corresponding threshold, which are set as 0.85 and 0.90 respectively.

4 Experimental Results

In this section, we will demonstrate some experimental results on real images of various contents, which are selected from the public resources of INRIA, CV datasets on the web, or our lab database. These images are under different imaging conditions, such as rotation, scale changes, viewpoint changes, etc.

Figure 3 shows show the different results for image pair Residence when one SIFT point or $N = 3$ SIFT points are used. Figure 4 shows the different results for image pair East South when one SIFT point or $N = 3$ SIFT points are used.

By adopting the SIFT information to provide structure information and local neighbor similarity measurement, we achieve highly reliable and accurate corner matching between two images. The reason can be concluded as follows: SIFT first provide robust scale and rotation change invariant information for each corner point, then correlation computation, which provides local neighbor similarity measurement, is used to enhance the performance. But we still need to note that there are some unmatched corner points, due to the insufficient SIFT features surrounding them and sharp local neighbor texture change.

The experimental results on real images of various contents show that our improved approach can effectively match two uncalibrated image corners against significant rotation and scale changes.

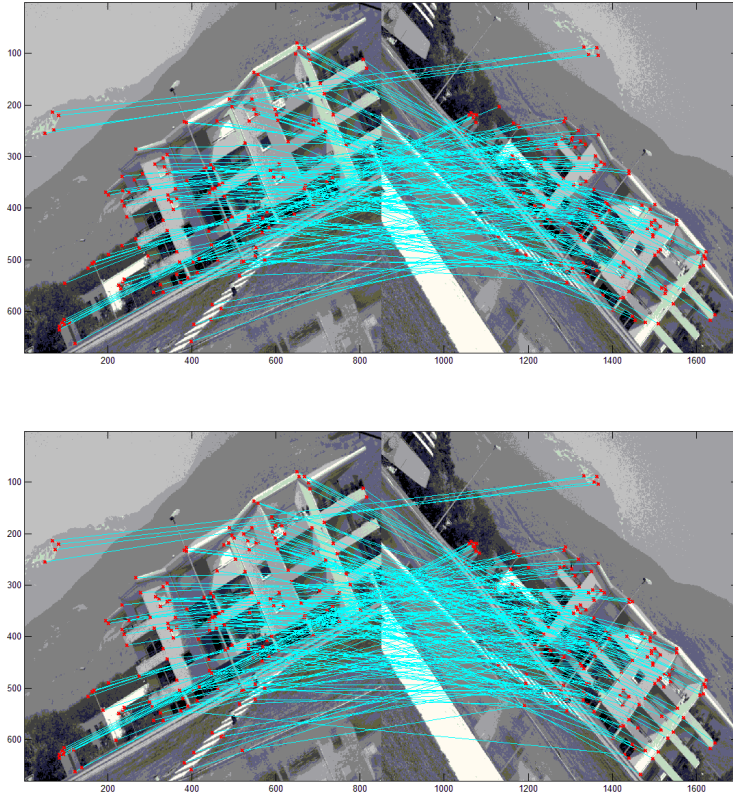


Fig. 3. Top: Matching result for the image pair Residence where only one SIFT point is utilized. Bottom: Matching result for the image pair Residence where $N = 3$ SIFT points are utilized.

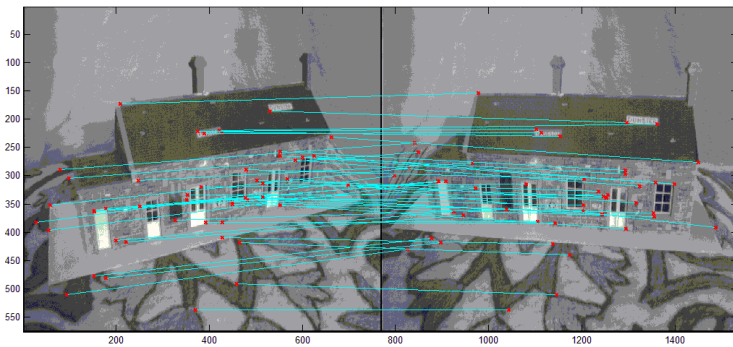


Fig. 4. Top: Matching result for the image pair House Model where only one SIFT point is utilized. Bottom: Matching result for the image pair House Model where $N = 3$ SIFT points are utilized.

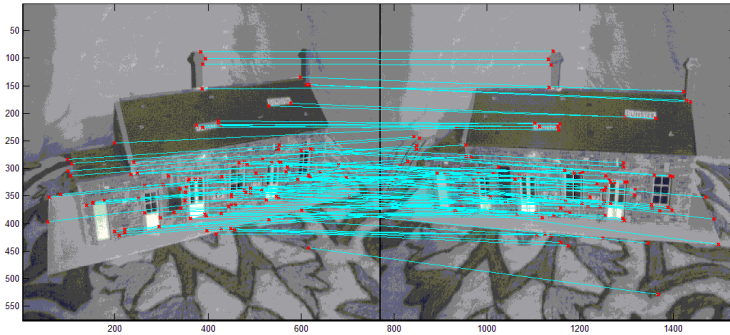


Fig. 4. (Continued)

5 Conclusion

In this paper, we propose an improved method for corner matching between two uncalibrated images, based both on blob information and local neighbor similarity between two interest corner points. Compared to traditional corner matching methods, our method can perform well when dealing with images with significant rotation and scale changes. Experimental results on real images verify the effectiveness and robustness of our method.

Acknowledgement. This work is supported in part by National Natural Science Foundation of China under Grant No. 61175014, Shanghai Pujiang Program under Grant No.11PJD023, and the Fundamental Research Funds for the Central Universities of China.

References

1. Zitová, B., Flusser, J.: Image registration methods: A survey. *Image and Vision Computing* 21, 997–1000 (1982)
2. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal* 78(1-2), 87–119 (1995)
3. Zhao, F., Huang, Q., Gao, W.: Image matching by normalized cross-correlation. In: *IEEE International Conference on Acoustics Speed and Signal Processing Proceedings* (2006)
4. Tuytelaars, T., Mikolajczyk, K.: Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision* 3(3), 177–280 (2007)
5. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple object class detection with a generative model. In: *Proc. of CVPR*, pp. 26–36 (2006)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: *ICCV*, pp. 1150–1157 (1999)

7. Shi, F., Huang, X., Duan, Y.: A Hybrid Approach for Robust Corner Matching. In: Tarn, T.-J., Chen, S.-B., Fang, G. (eds.) *Robotic Welding, Intelligence and Automation*. LNEE, vol. 88, pp. 169–177. Springer, Heidelberg (2011)
8. He, X.C., Yung, N.H.C.: Curvature scale space corner detector with adaptive threshold and dynamic region of support. In: *International Conference on Pattern Recognition*, pp. 791–794. IEEE Computer Society (2004)
9. Mokhtarian, F., Suomela, R.: Robust image corner detection through curvature scale space. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20, 1376–1381 (1998)
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: *4th Alvey Vision Conference*, pp. 147–151 (1988)
11. Pei, S.C., Ding, J.J.: Improved Harris' Algorithm For Corner And Edge Detections. In: *IEEE International Conference on Image Processing*, vol. 3, pp. 57–60 (2007)
12. Zhang, Y., Ji, D.: Adaptive Harris Corner Detection Algorithm Based on B-spline Function. In: *2nd International Conference on Intelligent Human-Machine Systems and Cybernetics*, pp. 69–72 (2010)

A Visual Dataflow Model for the Process Flow of Remote Sensing Products

Bing Zhou¹, Guan-feng Wu¹, Yong Xu¹, Jia-guo Li², and Yang Liu¹

¹ School of Computer and Information Engineering,
Henan University, Henan, 475004
zhoub1631@163.com

² Institute of Remote Sensing Applications,
Chinese Academy of Sciences, Beijing, 100101

Abstract. In order to conveniently and rapidly develop algorithms for remote sensing products, the basic idea is using some existing algorithms to develop a new algorithm. Due to the algorithm dependency, the algorithms are called one by one, which forms a process flow of remote sensing products. In this paper, a visual dataflow model is presented for the production of remote sensing products, which can represent the process flow of remote sensing products. The proposed model can reflect not only the relationship between algorithms, but also the number of algorithm to be called and the information of the data to be processed. Using this model, the changes of the process flow can be described conveniently and the concurrent execution of the algorithm can be performed.

Keywords: Data Flow Model, Visualization, Remote Sensing, Workflow Model.

1 Introduction

Along with the progress of the remote sensing technology, the applications of remote sensing expand quickly, such as weather forecasting, territory surveying, crop assessment, forest exploration, geological prospecting, environmental protection, disaster monitoring, urban planning, and so on. With the expansion of application fields, it raises a problem: how to develop a new remote sensing product, or remote sensing image processing algorithms conveniently and quickly?

A basic idea is to use some existing remote sensing processing algorithms as components to compose a new algorithm by the form of "building blocks". In other words, it makes the producing procedure of remote sensing products to be a process flow of some algorithm components.

This paper presents a visual dataflow model for remote sensing process flow, which could reflect not only the relationship between algorithms, but also the number of algorithm to be called and the information of the data to be processed in a visual way.

2 Related Work

The existing researches about automatically producing of remote sensing products or automatically processing of remote sensing images mainly use workflow model or flow chart to describe the process flow of remote sensing production[1,3,6]. Workflow model or flow chart can mainly reflect the dependence between algorithm components. But in some cases, they cannot reflect the relationship between algorithm components and remote sensing data intuitively and comprehensively. For example, it is not appropriate to describe the process flow of multi-period remote sensing products by workflow model or flow chart.

In this paper, a multi-period product means that producing such kind of products needs at least two remote sensing data of different period, comparing to single-period product, which only need one remote sensing data. The typical multi-period products include the product of Land Use Change Detection(LUCD) and the product of Gross Primary Productivity(GPP), etc. This paper uses GPP as an example.

There are lots of computational models and applications of GPP. Some need more than one period of NDVI (Normalized Difference Vegetation Index) to produce GPP[7-10]. Figure 1 shows a relatively simple process flow of GPP described by workflow model. Algorithm components are executed in accordance with the direction of the arrow. In figure 1, the block with double border means multi-period product, and the block with single border means single-period product.

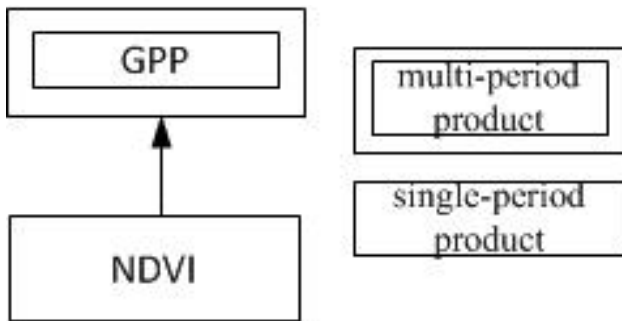


Fig. 1. The process flow of GPP described by workflow model

But in the real production procedure, because GPP is multi-period product, it needs at least two group data of different periods. It means that the NDVI will be called many times. But GPP is called only once. In this example, the workflow model neither reflects the calling times of NDVI, nor reflects the relationship between NDVI and source data.

If we consider more complex situation, in the case of many different periods of data and one NDVI product has already been produced for one period data, in order to reduce the time costs, the existing NDVI product can be used directly, rather than be repetitively produced again. It means that the NDVI algorithm component will not be called for the period, which has an existing NDVI product, but for the rest periods,

the NDVI algorithm component needs to be called. So it is very difficult for the workflow model to reflect this situation.

This paper proposes a dataflow model to describe the process flow of remote sensing production. In dataflow model, the data is the core and the processing is attached to the data. This dataflow model reflects that the nature of remote sensing production is the processing of data.

In addition, for better understanding of the entire process flow of remote sensing production, this paper gives a visual representation of the proposed dataflow model.

3 Visual Dataflow Model

The proposed visual dataflow model for the process flow of remote sensing products is composed of a number of visual nodes. A remote sensing process flow can be described through a combination of these visual nodes. These visual nodes are divided into the following categories:

- Dataflow node

The symbol of the dataflow node is \rightarrow . It means the direction that data goes from one side to the other. The side with an arrow called in-side. The other side is called out-side. Dataflow node can fork to different directions, because same data may be processed by different algorithm components.

- Data source node

Data source node represents the original remote sensing data. The visual symbol is \odot . This node can only be connected with out-side of one dataflow. Two data source nodes represent two different data sources. The dataflow deriving from different data source nodes are different.

- Data storage node

Data storage node represents persistent storage of dataflow, such as files, etc. The symbol of this kind of node is \blacksquare . The node can only be connected to the in-side of just one dataflow.

- Process node

Process node represents one process in the whole process flow of remote sensing products. The visual symbol is \square . The node is at least connected to one dataflow in-side, but can only connect to one dataflow out-side. If it is connected with two or more dataflow in-side, the dataflow must be different, which requires the dataflow derive from different data source nodes, or come from different process nodes.

The above four kinds of node can be named for certain specific meaning. If two nodes are named equally, these two nodes have same meaning. The named node

appears only once generally, but when necessary, it can occur many times (in order to facilitate the arrangement of layout).

The directed acyclic graph composed by above nodes interconnected together forms a visual dataflow model, which represents the process flow of remote sensing products. This means a visual dataflow model does not contain cyclic call between different algorithms.

4 The Application of Dataflow Model

Here NPP (Net Primary Productivity) is used as an example to elaborate the application of the visual dataflow model.

The producing procedure of NPP described by workflow model is shown in figure 2[7-12]. Starting from the bottom along the direction of the arrow, the algorithm components called in turn. Among them, the NPP, NDVI and TOAL (The Apparent Reflectance) are single-period products. Because GPP is a multi-period product, multiple sets of remote sensing data are needed. However, in figure 2, it only reflects what algorithm needs to be call, but misses any information about data and the calling times of algorithm.

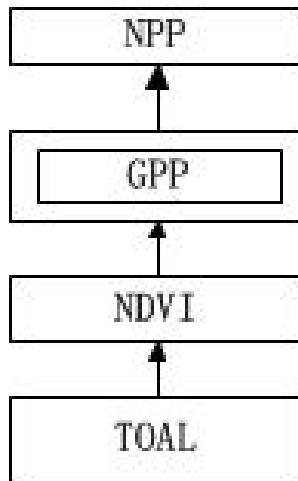


Fig. 2. The NPP workflow model

The process flow of NPP described by the proposed visual dataflow model is shown in figure 3, using two data source nodes for two different period remote sensing data. And then two TOAL and NDVI products are generated respectively. Because GPP is a multi-period product, two NDVI products are used to generate one GPP product and then one NPP product further.

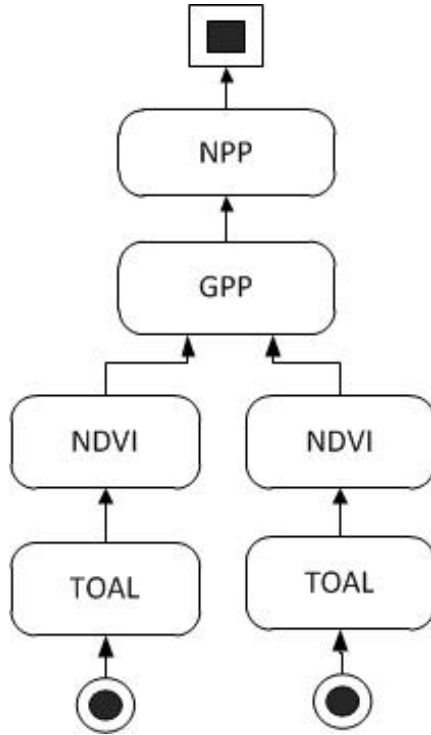


Fig. 3. The NPP Data flow model

Obviously, the model not only reflects what algorithm components are called, but also shows the calling times of algorithm and the information of the data to be processed.

5 The Changes of Process Flow Described by Dataflow Model

Sometimes it needs to change the process flow to skip some process nodes, because in some special situations some products have already been produced and need not to be re-generated. The proposed dataflow model can well describe the changes of process flow.

Suppose a complete dataflow model of a multi-period product is shown in figure 4. In this model all the products are single-period products except E, and two data source nodes are used.

Suppose D1 already exists. First, change D1 from process node to data source node. Then, delete the dataflow whose in-sides are connected with D1. Accordingly, the out-degree of C1 and B1, which connect with out-side of the deleted dataflow, should be reduced by 1. Because the out-degree of C1 and B1 become 0, these two nodes are deleted. That means C1 and B1 need not to be produced. Continue to repeat

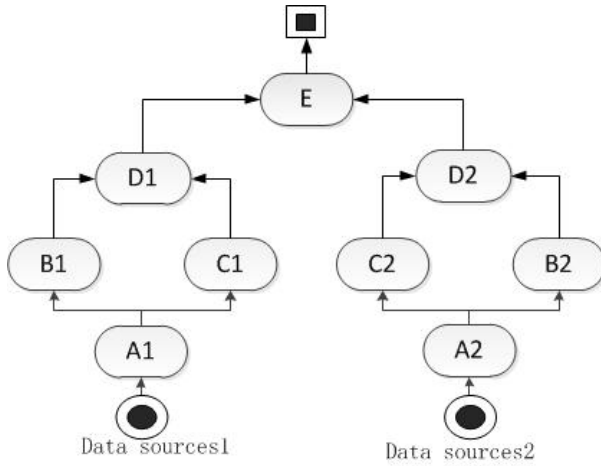


Fig. 4. An example of Dataflow model of a Multi-period product

similar steps, delete the dataflow which point to C1 and B1, the degree of the A1 become 0 after be reduced by 2; delete A1 and the dataflow which point to A1, the out-degree of data sources1 minus 1 to be 0, delete Data source1. The changed dataflow model is shown in figure 5 below.

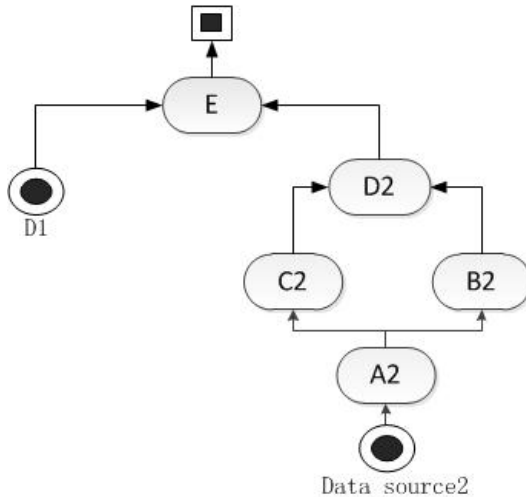


Fig. 5. The changed dataflow model

The changing operation of the dataflow model is:

1. Change the process node, which does not need to be produced, into a data source node.
2. Delete the dataflow that point to the deleted node.
3. Out-degree of the nodes, which connect with the deleted dataflow in step 2, minus 1.
4. If out-degree becomes 0, remove the nodes, and repeat (2) to (4).

6 Concurrent Execution of the Algorithm Components in the Dataflow Model

In order to improve the speed of production, if two process nodes do not have dependence relationship, they can be called parallel. Using the dataflow model can easily identify which process nodes can be concurrently executed.

Consider the following example as shown in figure 6, if adopting concurrent execution mechanism, the procedure can be divided into five steps. First, execute V4, V6, V8 and V9 concurrently. Second, execute V7. Third, execute V5. Fourth, concurrent execute V2 and V3. At last, execute V1 to get the finally product.

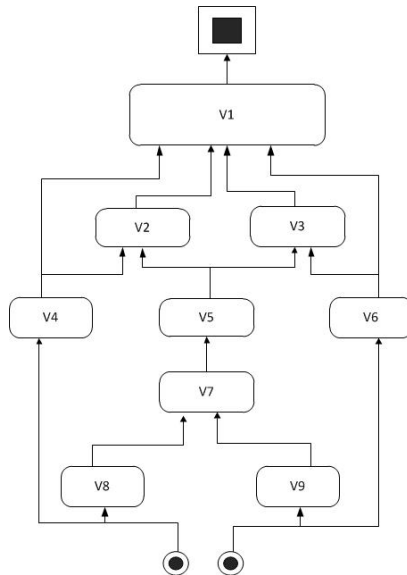


Fig. 6. The data flow model example

The concurrent execution strategy as follows:

1. Set up collection $A = \{V1, V2, \dots, Vn\}$, a collection of all un-called processing nodes. Set collection B as already been called process nodes. Set B is initialized to empty.
2. Initialize the in-degree of each process node in set A.
3. For all process nodes in the set A, if the in-degree equal 0, execute the corresponding algorithm, and move the node from set A to set B.
4. When the execution of a process node finishing, the in-degree of the nodes, which connect to the finished node, minus 1.
5. Continue (3) to (4), until all of the nodes in set A were removed to set B.

7 Conclusion and Future Work

Traditional workflow model or flow chart can only represent the dependence between algorithm components, but the proposed visual dataflow model can reflect the calling times of algorithms as well as the dependence between algorithm components.

In the case of many different period data, workflow model or flow chart cannot describe which data need to be used to produce which product. But using this model, it can easily describe the relationship between data and product.

Meanwhile our work states how to change a process flow and how to achieve concurrent execution within a visual dataflow model.

By far, this model can only describe the automatic process flow of remote sensing products. If there need manually operations, such as choosing training samples for land classification, this model cannot describe how to trigger the operation and when the operation finish.

We will improve this model continually, and other works includes developing a toolkit to draw and validate a visual dataflow model.

References

1. Xia, D.W., Wang, H., Shi, S.X.: Design on flow of visualization modeling system for ocean remote sensing information extraction. *Acta Oceanologica Sinica* 27, 97–103 (2005)
2. Wang, C.Y., Zhao, Z.M.: Design and implementation of processing flow system for remote sensing image. *Science of Surveying and Mapping* 31, 105–106 (2006)
3. Wei, J.X., Wei, D., Wu, X.C.: Workflow-based remote sensing visualization modeling system and its distributed scheduling algorithm. *Arid Land Geography* 32, 304–309 (2009)
4. Fan, J.T., Li, G.Q., Kang, L.: Data Storage in Work flow of Image Processing. *Computer Simulation* 21, 182–184 (2007)
5. Wang, R.R., Wang, L.Y.: An Event-Triggered Concurrent Dataflow Model. *Journal of Software* 14, 409–414 (2003)
6. Li, J.S., Chen, Y.W., Liu, D.S.: The Design and Implement of Multi-satellite Ground Pre-processing System Based on Workflow Technology. *Remote Sensing Technology and Application* 23, 428–433 (2008)
7. Liu, M.: Study on Estimation and Uncertainty of Terrestrial Ecosystem Productivity Based on RS and GIS—Take the Grassland Transect in Tibetan Plateau for Example. Liu Min. Nanjing normal university master's thesis (2008)
8. Zhao, J.J., Liu, L.Y., Xu, Z.W.: Monitoring winter wheat GPP in Huabei Plain using remote sensing and flux tower. *Transactions of the CSAE* 27, 346–351 (2011)
9. Prince, S.D., Goward, S.N.: Global Primary Production: a remote sensing approach. *Journal of Biogeography* 22, 815–835 (1995)
10. Huete, A.R., Liu, H.Q., Batehily: A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment* 59, 440–451 (1997)
11. Wu, Z.J., Xu, H.Q.: A New Index for Vegetation Enhancements of Mountainous Regions Based on Satellite Image Data. *Earth Information Science* 13, 656–664 (2011)
12. Chi, W.H., Zhou, G.S., Xu, Z.Z.: Apparent Reflectance and Its Applications in Vegetation Remote Sensing. *Acta Phytocologica Sinica* 29, 74–80 (2005)

Image Denoising Algorithm Based on Edge-Preserving Self-Snake Model and Wavelet-Based PDE

Changxiong Zhou¹, Shufen Lui¹, Tingqin Yan^{1,2}, and Wenlin Tao¹

¹ Department of Electronic and Informational Engineering, Suzhou Vocational University, Suzhou, Jiangsu, 215104, China

² Suzhou Key Lab. of Digital Design & Manufacturing Technology, Jiangsu 215104, China
handzhou@sina.com

Abstract. In this paper, a so-called edge-preserving self-snake model (EPSSM) that is able to remove noise while preserving edge features will be built, and using wavelet and partial differential equation (PDE), an integrated algorithm of wavelet-based PDE (IAWP) of image denoising is proposed. In integrated algorithm, The EPSSM is firstly used to remove noise of an input image, then we decompose the processed image by wavelet transform and its three high frequency coefficients are filtered by the EPSSM, finally, denoised image is reconstructed using inverse wavelet transform. The denoising performance of two proposed algorithms is measured according to PSNR values, and the experiments show that our methods have a better performance than others.

Keywords: image denoising, wavelet, partial differential equation, edge-preserving self-snake model.

1 Introduction

Image processing is an important area in the information industry. A crucial research is how to filter noise caused by the nature, system and processing of transfers and so on. In any image denoising algorithm, it is very important that the denoising process has less blurring effect on the image, and makes less changes or relocation to image edges. Denoising of image data has been an active area of research, with several different approaches being proposed using techniques such as wavelets [1,2], partial differential equation (PDE) [3,4], filtering with combination of wavelet and PDE [5,6], etc.

Wavelets have become the most important tool for image processing tools in recent years. It is an effective denoising tool in image processing. However for very noisy images, denoising using wavelets band pass filters or threshold may not work well since the edge features of the images may be lost while removing the noise. Since images are considered as non-stationary processes [7], adaptive filters are mostly applicable in image noise reduction. Some of these filters can be designed using partial differential equations which are employed as an efficient method for noise reduction in many studies. As well known, these equations such as non-linear, isotropic

diffusion and self-snake model are defined as the case where the diffusivity is a scalar function varying with the location in the image [3-5, 8-9]. The diffusivity varies with both the location and the orientation of the local image geometry, so these models can remove noise well and at the same time can protect the edge features and details of images.

In this paper, an edge-preserving self-snake model (EPSSM) and an integrated algorithm of wavelet-based PDE (IAWP) are applied for image denoising. In the proposed approach of IAWP, the EPSSM is initially applied on the input image. Then, processed image is decomposed using wavelet transform, the sub-bands containing the approximate coefficients are maintained, and the EPSSM is applied on detail coefficients. Finally, we reconstruct the image using inverse wavelet transform.

The remainder of this paper is structured as follows. In Section 2, we introduce the self-snake model. In section 3, the so-called edge-preserving self-snake model will be built and the integrated algorithm of wavelet-based PDE is proposed. In section 4, we analyze and discuss our algorithms by comparative results on noisy images. Finally, conclusions are drawn in Section 5.

2 Self-snake Model

The theory of image denoising based on PDE is often linked with heat conduction equation. Nonlinear diffusion PDE is a popular denoising approach in which prior information of image features can be incorporated via a diffusion coefficient into the denoising processing. Methods based on nonlinear diffusion approaches have been applied for a variety of applications [3-6].

Perona and Malik firstly proposed a nonlinear anisotropic diffusion equation, which basic idea can be attributed to the initial value of the input image to solve the nonlinear diffusion equation. We can also define energy function via

$$E = \int_0^1 g(\nabla(u(s))) ds \quad . \quad (1)$$

Where ∇ is the gradient operator, u is gray value of image and g is an edge detector defined by a positive decreasing function. The snake model is to minimize E of above energy function, using the variational level set formulation, we arrive at

$$\partial u / \partial t = \text{div} (g(|\nabla u|) \nabla u / |\nabla u|) |\nabla u| \quad . \quad (2)$$

Where div is the divergence operator, Eq.(2) is named self-snake[9], so we may write

$$\partial u / \partial t = g(|\nabla u|) \text{div} (\nabla u / |\nabla u|) |\nabla u| + \nabla g(|\nabla u|) \cdot \nabla u \quad . \quad (3)$$

The first item is called direction diffusion with an edge stop function which can reject diffusion at edges and permit smoothing in other places, and second item is named shock filter which sharpens edges efficiently. Where stop function $g(x)$ can be defined as follows:

$$g(|\nabla u|) = 1 / ((|\nabla u| / k)^2 + 1) \tag{4}$$

This approach is simple but if edges are partly lost to a certain extent or are contaminated severely by noise, it may not be able to recover them and thus fail to preserve other features that are bounded by the edges.

3 Our Proposed Method

3.1 Edge-Preserving Self-Snake Model (EPSSM)

By nonlocal derivative (NLD) the grayscale difference between two pixels is measured by two regions (patches) centred at the pixels, the new feature detector which measures image intensity contrasts between neighboring patches in a more sophisticated manner and can effectively capture fundamental edge features [10].

We describe the concept of the nonlocal difference in a one-dimensional (1-D) image. Extension to the two-dimensional (2-D) case is straightforward and will be discussed later. Let $I : \Omega \rightarrow R^1$ be a 1-D scalar image defined on the discrete domain Ω and $x \in \Omega$ is the pixel position, $x = [x_1, x_2, \dots, x_i, \dots, x_N]$. For each pixel x , we define a neighborhood region N_x , which comprises W pixels centred around x . We further define a patch P_x , which is a vector comprising gray-level value of all pixels within the neighborhood region N_x .

$$P_{x_i} = [x_{i-(W-1)/2}, x_{i-(W-1)/2+1}, \dots, x_i, \dots, x_{i+(W-1)/2}] \tag{5}$$

here W is assumed to be an odd number for symmetry consideration. The nonlocal difference between two signal samples $I(x_i)$ and $I(x_j)$ can be measured as a Gaussian weighted Euclidean difference,

$$d_{NL}(x_i, x_j) = \left\| P_{x_i} - P_{x_j} \right\|_{2,\sigma} \tag{6}$$

of two vectors P and in a W -dimensional space, where σ is the standard deviation (Std) of a normalized Gaussian kernel.

$$|\nabla_{NL} x_i| = \left\| P_{x_{i-(W+1)/2}} - P_{x_{i+(W+1)/2}} \right\|_{2,\sigma} = \left(\sum_{k=-(W-1)/2}^{k=(W-1)/2} G_\sigma(k) (x_{i-(W+1)/2+k} - x_{i+(W+1)/2+k})^2 \right)^{1/2} \tag{7}$$

From Eq.(6) we can define a nonlocal derivative at pixel x_i when x_j approaches x_i . Where G_σ is a Gaussian kernel with Std σ . For two-dimensional image $u(i, j)$,

we may define that the nonlocal derivative $|\nabla_{NL} u(i, j)|$ at pixel (i, j) is the corresponding maximum value in horizontal, vertical and both diagonal directions,

$$|\nabla_{NL} u(i, j)| = \max(|\nabla_{NL} u(i, j)|_{\text{horizontal, vertical and both diagonal directions}}). \tag{8}$$

The nonlocal derivative involves the difference between two adjacent patches. For the same reason, Eq. (7) and Eq. (8) are more reliable than the pixel-level gradient operator involving two pixels to measure edges under noise contamination. In Eq. (4), if the gradient operator ∇ of stop function is replaced by the nonlocal derivative, then Eq. (2) becomes Eq. (9) as follow:

$$\partial u / \partial t = \text{div} (g (|\nabla_{NL} u|) \nabla u / |\nabla u|) \nabla u \tag{9}$$

Above Eq. is named edge-preserving self-snake model (EPSSM) which is reliable than self-snake model under noise contamination, so the EPSSM can remove noise of images while preserving edge features. The numerical solution of the EPSSM is described below. Let the time step be τ and the spatial step be h in x, y directions, then the time and space coordinates can be described as:

$$\begin{aligned} t &= n\tau, n = 0, 1, 2, \dots; x = ih, y = jh, \\ i &= 0, 1, 2, \dots, M - 1, j = 0, 1, 2, \dots, N - 1 \end{aligned} \tag{10}$$

where $M \times N$ is the size of the image. Let $u_{i,j}^n = u(ih, jh, n\Delta)$, and

$$v_{i,j} = \text{div} (g (|\nabla_{NL} u(i, j)|) \nabla u / |\nabla u|)_{i,j} \tag{11}$$

Equation (9) belongs to Hamilton-Jacobi equations and needs upwind scheme, so the explicit numerical scheme as following:

$$u_{i,j}^{n+1} = u_{i,j}^n + \tau (\max(v_{i,j}, 0) \nabla_{i,j}^- + \min(v_{i,j}, 0) \nabla_{i,j}^+) \tag{12}$$

Where the time step τ should satisfy CFL condition, experimental results show that iteration number n should be more than 5 times.

3.2 Integrated Algorithm of Wavelet-Based PDE (IAWP)

In recent years, many researchers have performed wavelet analysis for image denoising. Two general approaches are commonly used in setting the wavelet coefficients to zero. The first one is low pass filter, since the high frequency part essentially corresponds to the noise; one set all coefficients of these modes to zero. Unfortunately, this approach can sometimes destroy edges features of the image as these features are also in the high frequency range. The second approach is thresholding, which is a simple non-linear technique operating on the wavelet transform coefficients. If the coefficient is smaller than the threshold, it is set to zero; otherwise it is kept or modified as defined in the used thresholding method. But for very noisy images, two wavelet transform approaches cannot keep edge features as well as removing noise.

The wavelet analysis and PDE are represented as a mathematical tool in image processing in various fields. In order to comprehensive advantages of wavelet and the PDE for image denoising, overcome their on shortcomings, using PDE represented by EPSSM and the combination of wavelet transform, this paper proposed a integrated algorithm of wavelet-based PDE, which is shown in Figure 1.

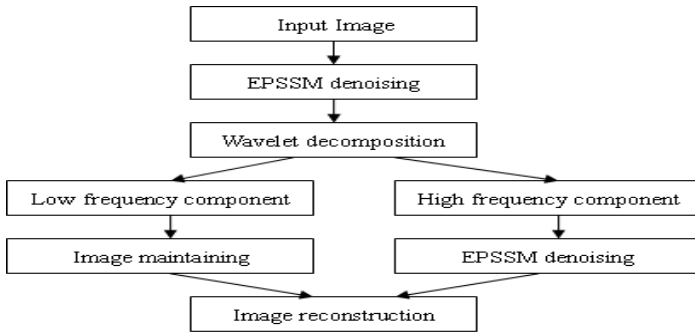


Fig. 1. Structure of integrated algorithm of wavelet-based PDE

The steps of the proposed integrated algorithm are:

Stage I: for an input image, we use PDE represented by edge-preserving self-snake model to remove noise while preserving edge features;

Stage II: computing the wavelet decomposition of output image of step one, the image is split into a low frequency component and three high frequency components;

Stage III: while the low frequency component is maintained, three high frequency components are filtered by EPSSM for denoising;

Final Stage: Applying the inverse wavelet transform, a reconstructed image with lower noise is obtained.

4 Experiments and Comparisons

This section aims at demonstrating the denoising performance of EPSSM and IAWP respectively. We show the experiment results with four methods of the wavelet thresholding, the method in Ref. [5], the proposed EPSSM and IAWP in various noise levels. All of the experiments were performed on images of LENA and CAMERA. We study the PSNR performance of the four methods using images corrupted with Gaussian noises (mean = 0 and variance = 0.01 to 0.04), salt & pepper (density=0.01 to 0.04) noises, and combination of Gaussian noises (mean = 0 and variance = 0.01 to 0.04) and salt & pepper (density= 0.01 to 0.04) noises respectively.

Table 1 shows the PSNR values of four methods for the images of Lena corrupted with Gaussian noises. Data in the table 1 shows that the PSNR values of the image denoised with EPSSM and IAWP are higher than that with the wavelet thresholding and the method of Ref. [5] in different noise levels. The EPSSM involves the nonlocal derivative between two adjacent patches and is more reliable than other methods to

Table 1. Comparison of PSNR for Lena corrupted with Gaussian noises by four methods

Noise variance	Noisy image	Wavelet thresholding	Ref.[5]	EPSSM	IAWP
0.01	20.08	22.24	22.90	27.51	27.41
0.02	17.22	19.49	20.42	25.13	25.61
0.03	15.57	17.90	18.41	23.41	24.19
0.04	14.49	16.85	17.33	22.30	23.34

measure edges under noise contamination, so it gains 5 dB higher PSNR than the method of Ref. [5]. Experimental results also shows that the IAWP is the best one in four denoising methods when noise level is more than 0.02, since it is an integrated algorithm which has advantages of wavelet and the PDE for image denoising.

Table 2 shows the PSNR values of four methods for the images of Lena corrupted with salt and pepper noises. Data in the table 2 shows that the PSNR values of the image denoised with EPSSM and IAWP are higher than that with the wavelet thresholding and the method of Ref. [5] in different noise density levels, and shows that IAWP's performance is inferior to that of the EPSSM when noise density level is less than 0.03, since processed image is over filtered by the IAWP.

Table 2. Comparison of PSNR for Lena corrupted with salt & pepper noises by four methods

Noise density	Noisy image	Wavelet thresholding	Ref.[5]	EPSSM	IAWP
0.01	25.57	25.95	28.19	29.70	28.90
0.02	22.38	23.15	25.57	28.48	28.07
0.03	20.44	21.52	23.71	27.31	27.25
0.04	19.56	20.80	22.54	26.73	26.82

Table 3. shows the PSNR values of four methods for the images of Lena corrupted with combination of Gaussian noises and salt & pepper noises.

Table 3. Comparison of SNR for Lena corrupted with combination noises by four methods

Noise variance & density	Noisy image	Wavelet thresholding	Ref.[5]	EPSSM	IAWP
0.01	19.04	21.07	22.16	26.64	26.76
0.02	16.12	18.28	19.14	23.99	24.68
0.03	14.52	16.82	17.33	22.15	23.14
0.04	13.36	15.72	16.16	20.82	21.97

Data in the table 3 shows that the PSNR values of the image denoised with EPSSM and IAWP are higher than that with the wavelet thresholding and the method of Ref. [5] in different noise levels, and shows that IAWP's performance is prior to the EPSSM at every noise level. Since the IAWP has advantages of wavelet and the PDE, its performance is the best one of four methods for noisy image which is contaminated severely.

Experiment results on Camera are shown as Fig.2. Fig.2 (a) is clear image of Camera Lena. Fig.2 (b) shows Camera after being corrupted with a combination of Gaussian noises (mean = 0 and variance = 0.03) and salt & pepper (density= 0.03) noises respectively. Fig.2 (c) shows Camera after filtering with wavelet thresholding. Fig.2 (d) shows Camera after filtering with the method in Ref. [5]. Fig.2 (e) shows Camera after filtering with EPSSM. Fig.2 (f) shows Camera after filtering with IAWP. From visual sense effect of human eyes, Fig.2 shows that two proposed approaches of EPSSM and IAWP have good denoising performance for Camera compared with the wavelet thresholding and the method in Ref. [5]. Two proposed approaches not only remove the noise well but also reserve the edge and details information in the image.

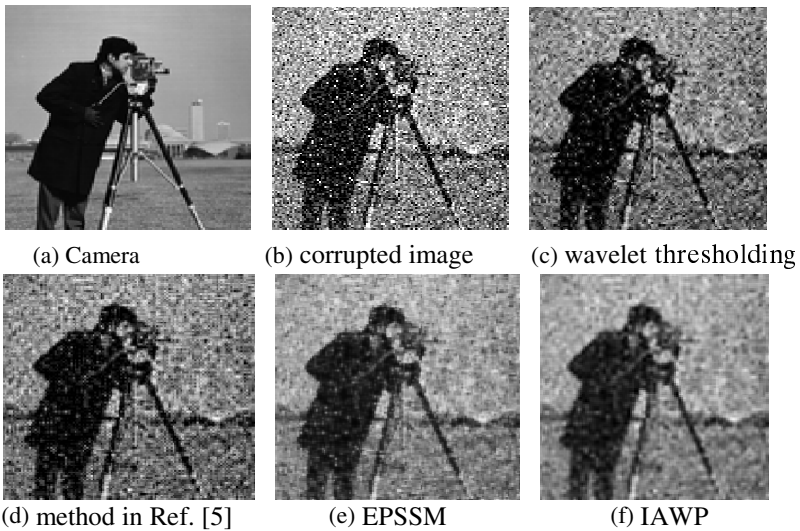


Fig. 2. Noisy image of Camera (Gaussian noises variance = 0.03 and salt & pepper density = 0.03) and results of denoising by four methods

5 Conclusions

In this paper, we introduce nonlocal derivative and construct an edge-preserving self-snake model (EPSSM) which can reject diffusion at edges and permit smoothing in other places, so remove noise of images while preserving edge features, as well as give the numerical scheme of the EPSSM. Simulation experiments show that it is prior to the wavelet thresholding and the method of Ref. [5]. Furthermore, an integrated algorithm of wavelet-based PDE (IAWP) is proposed. In wavelet transform, the sub-band containing the approximate coefficient is maintained, and EPSSM is applied on detail coefficients, so IAWP is prior to the EPSSM for noisy image which is contaminated severely by combination of Gaussian noises and salt & pepper noises.

Acknowledgements. This research was sponsored by the grants of Natural Science Foundation of China (No. 60970058), the grants of Natural Science Foundation of Jiangsu Province of China (No.BK2009131), Innovative Team Foundation of Suzhou Vocational University (No.3100125), Suzhou Infrastructure Construction Project of Science and Technology (No.SZS201009) and Qing Lan Project of Jiangsu Province of China.

References

1. Shui, P.L., Zhou, Z.F., Li, J.X.: Image Denoising Algorithm via Best Wavelet Packet Base Using Wiener Cost Function. *IET Image Process* 1(3), 311–318 (2007)
2. Zhang, W., Yu, F., Guo, H.M.: Improved Adaptive Wavelet Threshold for Image Denoising. In: 2009 Chinese Control and Decision Conference, pp. 5958–5963 (2009)
3. Perona, P., Malik, J.: Scale Space And Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(7), 629–639 (1990)
4. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Image Enhancement And Denoising by Complex Diffusion Process. *IEEE Trans. Patt. Anal. Mach. Intell.* 26(8), 1020–1036 (2004)
5. Cao, Z.W., Zhang, X.Z.: PDE-Based Non-Linear Anisotropic Diffusion Techniques for Medical Image Denoising. In: 2012 Spring World Congress on Engineering and Technology, Xi'an, China, May 27-30 (2012), doi:10.1109/SCET.2012.6341990
6. Nadernejad, E., Forchhammer, S.: Wavelet-Based Image Enhancement Using Fourth Order PDE. In: *IEEE International Symposium on Intelligent Signal Processing*, pp. 116–121 (2011)
7. Azzabou, N., Paragios, N., Guichard, F.: Uniform and textured regions separation in natural images towards MPM adaptive denoising. In: Sgallari, F., Murli, A., Paragios, N. (eds.) *SSVM 2007*. LNCS, vol. 4485, pp. 418–429. Springer, Heidelberg (2007)
8. Blanc-Feraud, L., Charbonnier, P., Aubert, G., Barlaud, M.: Non-Linear Image Processing: Modeling And Fast Algorithm for Regularization with Edge Detection. In: *Proceedings of the IEEE International Conference on Image Processing*, vol. 1, pp. 474–477 (1995)
9. Xu, C., Tang, W.: A Wavelet And PDE Method for Optical Image Interpolation. In: 2nd International Conference on Information Science and Engineering, Hangzhou, China, pp. 3498–3501 (2010)
10. Qiu, Z., Yang, L., Lu, W.P.: A New Feature-Preserving Nonlinear Anisotropic Diffusion for Denoising Images Containing Blobs And Ridges. *Pattern Recognition Letters* 33(3), 319–330 (2012)

Single Sample Face Recognition Based on Multiple Features and Twice Classification

Xiaohua Wang^{1,2}, Wei Liu^{1,2}, Min Hu^{1,2}, and Liangfeng Xu¹

¹ School of Computer and Information, Hefei University of Technology

Xh_wang@hfut.edu.cn

² Affective Computing and Advanced Intelligent Machines Anhui Key Laboratory

yimilwwl@163.com

Abstract. In order to improve the performance of face recognition with single sample effectively, a face recognition method based on multiple features and twice classification is proposed. For obtaining sufficient face information, facial multiple features combining differential excitation and Compound Local Binary Pattern (CLBP) on the asymmetric region are extracted. Elastic Matching (EM) has better robustness for pose. However, the computation complexity of the method is rather high. Classifying twice strategy is proposed to short the time of data processing. Experimental results on ORL database and FERET database show that the method is effective in getting better recognition rate and speed, also has a certain robustness to pose.

Keywords: Single sample face recognition, differential excitation, compound local binary pattern (CLBP), elastic matching (EM).

1 Introduction

Face recognition has been extensively applied in many fields including identification, video retrieval and security monitoring et al. Face recognition has become a research hotspot with great progress. Many algorithms like Gabor wavelet, Local Binary Pattern (LBP) have been used to face recognition by many researchers and achieved better results[1-2]. But many studies focus on a number of training samples [1-4]. On some special occasions such as ID verification, Passport verification, we only can get one image. Hence, single sample recognition becomes very important. There are many face variations (pose, illumination, facial expression and partial occlusion) in real-life. Sometimes the differences of face images from one person under different conditions are greater than that of different persons under the same condition. Obviously, one image just offers numbered face information. Therefore, correct identification with one training sample has become a difficult problem.

The rest of this paper is organized as follows. Related works is given in section 2. In section 3, we describe the details of multiple features. The process of twice classification is illustrated in section 4. Section 5 presents the experiments on ORL and FERET face database. Finally, conclusions is shown in Section 6.

2 Related Work

Many methods have been proposed currently to solve single sample problem and can be classified as sample expansion method [5-6], 3D recognition method [7], feature subspace expansion method [8-9], image enhancement method [10] and so on. Zhu and Shan expand one training sample into many virtual images based on certain criterions [5-6]. But they are highly close with the original sample. As a result, their method has not a high recognition rate. What's more, based on the research of 3D on one single sample problem is valid and effective compared with 2D [7]. However, the large amount of data and computation are unaccepted, and real-time performance is low efficiency. If traditional methods of characteristic subspace are used for single sample face recognition directly, high accuracy could not be gotten easily. Thus Li improved (2D) 2 PCA and W2DPCA, which has been applied to single sample face recognition [9]. Wang proposed LBP features on different scales by gold tower to achieve more face information, but it does not consider the effect of pose [10].

As we know, differential excitation [11] can extract local salient features, and Compound Local Binary Pattern (CLBP) [12] can achieve subtle texture features, which have a complement to differential excitation features. Elastic matching (EM) through matching the best image block in a certain area to solve the displacement problem, so as to overcome the pose problem effectively. However, large matching image blocks will lead to a great amount of calculation and affect real-time of system.

Based on the above analysis, in order to extract more face information to get correct identification and the recognition speed at the same time is not affected, a face recognition method adopting multiple texture features and classifying twice is proposed. Firstly, differential excitation and CLBP on the asymmetric region [13] are combined to extract facial texture features. Then classify face twice. In the first classification, the Chi-square distance is used to find several candidates' faces, and the most similar face is gotten using EM to complete the second classification.

3 Multiple Features

Differential excitation is a part of WLD, which is inspired by Weber's Law. The law points out that the rate of an excitation changes relative to the points around is a constant. It can be described as:

$$\frac{\Delta I}{I} = K \quad (1)$$

where ΔI is the increment threshold, and K is a constant. I represents the initial intensity, which amount to the center pixel in image. ΔI is computed as:

$$\Delta I = \sum_{i=0}^{p-1} (\Delta x_i) = \sum_{i=0}^{p-1} (x_i - x_c) \quad (2)$$

ΔI can be considered as the sum of difference values between the center pixel x_c and its each neighbor pixel x_i .

The differential excitation only extracts facial salient texture features, but just one training sample's features can't contain enough face identifying information. LBP [14] can extract facial fine texture features quickly and efficiently. And it exists some complementary relationship between differential excitation features and LBP features. So this paper uses improved LBP operator to extract facial another feature.

Classical LBP operator just obtains the sign of the difference between the central pixel and its neighbor pixels. Faisal Ahmed et al. proposed CLBP, which contains two parts about sign and magnitude of the difference between the central pixel and the neighbor pixels, CLBP can be expressed as :

$$S(i_p - i_c) = \begin{cases} 00 & i_p < i_c, \quad |i_p - i_c| \leq u \\ 01 & i_p < i_c, \quad |i_p - i_c| \geq u \\ 10 & i_p \geq i_c, \quad |i_p - i_c| \leq u \\ 11 & i_p \geq i_c, \quad |i_p - i_c| \geq u \end{cases} \quad (3)$$

Here, i_c is the value of center pixel, i_p is the value of its neighbor pixel p. u represents the average value difference between the center pixel and the neighbor pixels.

The differential excitation map and CLBP map are shown in Fig. 1. We can see, the differential excitation contains local salient features of the image, fully reflects facial contour information. The CLBP map shows much subtle texture information, makes up for the defect of differential excitation.



Fig. 1. Differential excitation map and CLBP map

As classical LBP operator only achieves information on a small area (3x3 pixels) and circle operator omits several pixels inside circle. For the sake of obtaining the face information on a larger area and calculating every pixel of the area, we calculate multiple features on asymmetric region. You can see the detail of calculation on asymmetric region in literature 13. Next we calculate Asymmetric Region Compound Local Binary Pattern (ARCLBP) features, which is shown in Fig. 2. CLBP code will be divided into two sub-sections: the value (code is 00101001) of 1, 3, 5, 7 sub-region is 41; then the second sub-code 01111110 is gotten from 1,3,5,and 7 sub-region, which value is 126. According to the formula (1), (2), the value of Asymmetric Region Differential Excitation(ARDE) is obtained.

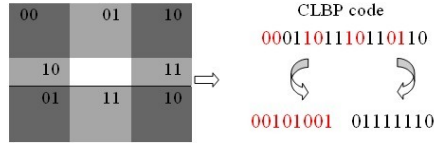


Fig. 2. Calculation process of CLBP

ARCLBP map and ARDE map are divided into 16 blocks separately. After that, for every map, all image blocks' histogram features are cascaded. At last multiple features are extracted by concatenating the histogram features of all maps, which can be as a feature vector. The process is illustrated in Fig. 3.

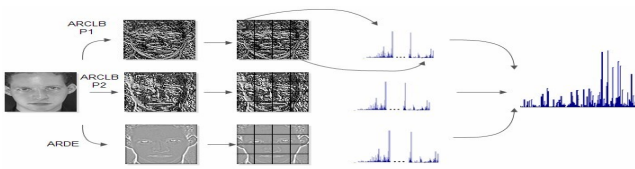


Fig. 3. The process of extracting multiple features

4 Classifying Twice

Due to the influence of pose, some parts of the human face appear some offset, which have a great impact on single sample face recognition. As Figure 4 Shown, the two image blocks in the same position of the image are marked as A and B (white box labeled), the area B is not corresponds to area A because of pose change. While the area C obtained by panning (black box labeled) is corresponding to the block A.

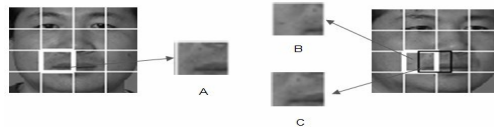


Fig. 4. Frontal face and rotated face

EM can reduce mismatching by finding the best matched block on the face search area of each image block to be matched. But if the test face is matched with each training sample, computation time will be long. This paper presents twice classification method to speed up recognition. Shorten the time of calculation of EM through reducing the number of training samples.

In the first classification, we measure feature distances between the test face and every face in training library using Chi-square distance, then sort faces by distance which is ascending. At last the top p faces are selected .

In the second phase, the test face is matched with these p faces through EM, and the final face is obtained by nearest neighbor classifier. Assuming the size of the

image block is $m \times n$ pixels, the search field is $M \times N$ pixels and the translations along the x-axis and y-axis both are L pixels. The steps are as follows:

Step 1. The distances between each image block of train samples and its matched blocks are calculated. $d_{j,k}$ ($k=1,\dots,b$) represents the distance between the k -th matched block and the j -th block of a candidate's face, which can be obtained using Chi square distance. Here, b is the number of matched blocks of each block of the p train samples. b is described as:

$$b = \left(\frac{M - m}{L} + 1 \right) \times \left(\frac{N - n}{L} + 1 \right) \tag{4}$$

Step 2. For each image block of train samples, we search the best matched block in its searching field. The distance between the j -th block of a candidate's face and its best matched block is

$$d_j = \min \{ d_{j,1}, \dots, d_{j,k}, \dots, d_{j,b} \} \tag{5}$$

Step 3. The total distance D between the candidate face and the test face:

$$D = \sum_{j=1}^B d_j \tag{6}$$

Where B is the total blocks per image. The face whose D is the smallest is chosen as the recognition face on the basis of nearest neighbor classifier.

The flowchart of face recognition is shown in Fig. 5.

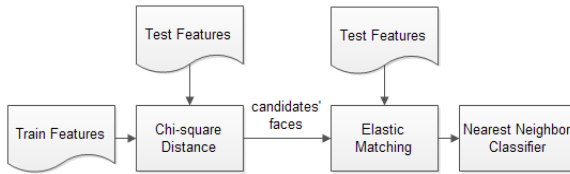


Fig. 5. Flowchart of face recognition

5 Experiments

The algorithm proposed is evaluated on ORL database and FERET database. ORL database includes 10 different images of 40 persons. Select randomly one image each person, total 40 images as a class, all can be divided into 10 classes. One class is chosen as training set, and other classes as test set. The process is repeated 10 times then average recognition rate is obtained. FERET face database comprises images of ordinary face, face with expression and pose. Experimental images are 1164 images, which consist of 194 persons and 6 images per person. The ordinary facial images of 194 persons are selected as training samples and 970 images remaining as test samples, we divide the test samples into pose set, expression sets and lighting set. Some images are shown in Fig. 6.

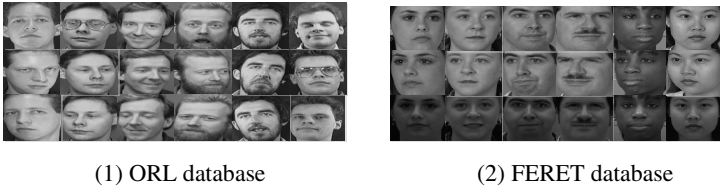


Fig. 6. Some sample images on ORL database and FERET database

In the experiments, all images have been normalized to 80×80 pixels and processed by histogram equalization. The size of image block is 20×20 pixels, and the search area of the image block is 40×40 pixels. The translations is set as 5 pixels.

5.1 Effect of Different p on Recognition

If p is too small, the correct face may be excluded in the first classification. However, if p is bigger, the speed of the second classification is slower. So the appropriate value of p is important. From Fig. 7 we can observe that the result attains a maximum while p is 5. When $p > 5$, recognition rate starts to level off, but the time is sustained growth. By the overall consideration of speed and accuracy, the value of p is set as 5.

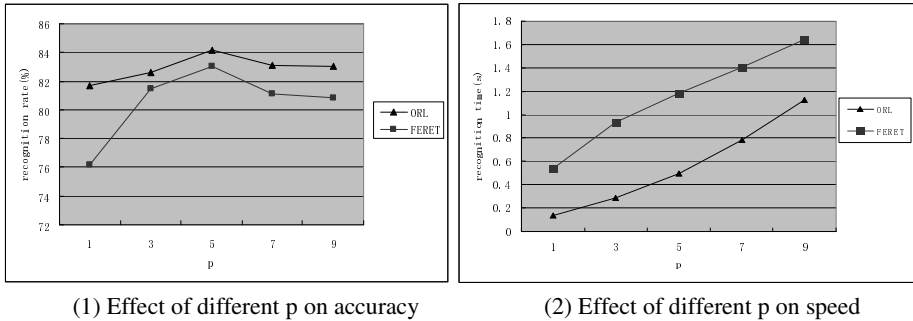


Fig. 7. Effect of different p on the accuracy and speed of face recognition with single sample

5.2 Effect of Multiple Features and Twice Classification on Recognition

For verifying the impact of the multiple features and the new classified method to system performance, we will give the experimental results of the following algorithms:

1. Based on ARCLBP histogram features method (AC);
2. Based on the method of the ARDE histogram features (AD);
3. Based on ARCLBP and ARDE histogram features method (ACAD);
4. Based on multiple histogram features and using EM directly (ACADE).

Table 1. Recognition rate and speed using ORL database

Algorithm	AC	AD	ACAD	ACADE	Our method
Group 1(%)	80.28	69.44	84.17	85.56	86.67
Group 2(%)	80.83	72.5	81.67	84.72	84.72
Group 3(%)	82.5	73.33	85	85.83	85.83
Group 4(%)	78.06	75.28	81.39	82.5	83.61
Group 5(%)	80.83	68.61	80.28	81.39	81.11
Group 6(%)	78.61	69.17	78.89	84.17	82.22
Group 7(%)	79.17	74.17	82.22	84.17	84.17
Group 8(%)	79.17	70.28	80	85.28	85.56
Group 9(%)	78.89	69.44	79.44	82.22	81.94
Group10(%)	79.44	76.11	84.17	84.72	86.11
Average recognition Rate(%)	79.78	71.83	81.72	84.06	84.19
Recognition Time(s)	0.109	0.081	0.137	5.030	0.496

As can be seen from the results in Table 1, the low recognition rate of AD reveals that AD features lack much face information. The recognition rate of AC and AD are lower than a combination of them about 1.94%, 9.89%, which indicate multiple texture features precede AD and AC. As EM is robustness for pose, both of ACADE and the proposed algorithm are surpass 84% at the recognition rate yet recognition in the proposed algorithm is slightly higher. The reason is that our method only matches with several similar faces and excludes dissimilar faces. However, the test face needs matching with every training face using ACADE, thereby recognition speed is very slow, reaches 5.03s per sample. Compared with other algorithms, the proposed algorithm have the highest average recognition rate and faster recognition speed.

Table 2. Recognition rate using FERET database

Algorithm	AC	AD	ACAD	ACADE	Our method
Pose Set 1(%)	75.26	62.89	76.80	81.96	85.57
Pose Set 2(%)	90.72	70.10	92.27	93.81	94.85
Pose Set 3(%)	49.48	24.23	53.61	58.76	63.92
Expression Set(%)	92.27	82.47	94.85	92.78	94.85
Lighting Set(%)	50.00	39.18	63.40	63.92	75.77
Average Recognition Rate(%)	71.55	55.78	76.19	78.25	83
Recognition Time(s)	0.455	0.353	0.539	23.045	1.180

As shown in Table 2, set 3's facial deflection angle is large relative to pose set 1, 2. The effect of multiple features is still better than single features, which reflect obviously the advantages of multiple features. The recognition rate of pose set have increased after adding EM. Differ from the results on ORL database, the proposed method is higher at about 4.75% than ACADE. Because twice classification method

can exclude more dissimilar faces on FERET database. With the number of training persons increasing, the time of ACADE increases accordingly, but our method changes little. It shows the superiority of proposed method on big face database.

5.3 Comparing Proposed Algorithm with Other Algorithms

LBP is a kind of effective face representation method, it is fit for single sample face recognition because of LBP features extraction just bearing on the sample itself. $WM(2D)^2PCA$ is obtained through combining partitioning and weight with $(2D)^2PCA$, which can efficaciously extract facial local features. The proposed algorithm is compared with $WM(2D)^2PCA$ and classical LBP in this section. The results are shown in Fig. 8. As our method extracts more facial features and considers pose variations within a narrow range, its effect is better than two other methods.

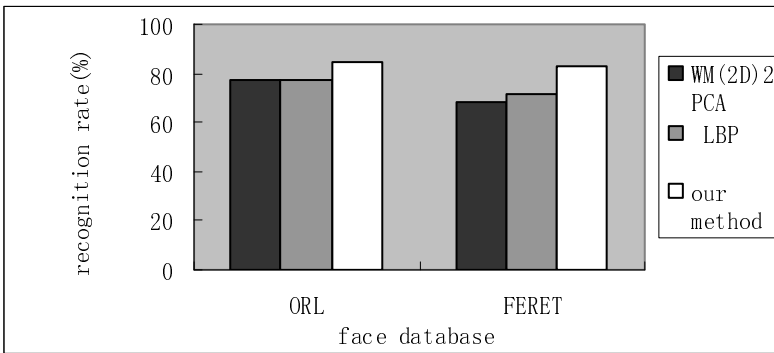


Fig. 8. Recognition rates of different methods

6 Conclusions

In this paper, a single sample face recognition method based on multiple features and twice classification is proposed. To a certain extent, the multiple features extract ample texture information. Moreover, EM algorithm is applied in classification to reduce pose interference and faces are classified twice, which improve the performance of single sample face recognition effectively. The advantages of the proposed method are proved on the experiments using ORL and FERT databases. Although the recognition rate of proposed method improves a little relative to ACADE algorithm, its speed far lower. And the superiority of our method is more apparent compared with other methods.

This paper doesn't consider the difference in rate of contribution of AD and AC. And it only extracts texture features without considering other type of features. In addition, the performance is not very good when pose changes greatly. These problems will be performed to solve in our following researches.

Acknowledgments. This work was supported in part by the National High-Tech Research & Development Program of China under Grant No.2012AA011103), Key Scientific and Technological Project of Anhui Province under Grant No 1206c0805039 and by the NSFC-Guandong Joint Foundation Key Project under Grant No. U1135003.

References

1. Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P., Konen, W.: Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers* 42(3), 300–311 (1993)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12), 2037–2041 (2006)
3. Fan, C.N., Zhang, F.Y.: Illumination Invariant Extraction on Gabor Phase. *Journal of Image and Graphics* 17(5), 676–681 (2012)
4. Zhang, J.W., Wang, W.L., Yao, X.M., Shi, H.Y.: Face Recognition Using Tensor Local Fisher Discriminant Analysis. *Acta Automatica Sinica* 38(9), 1485–1495 (2012)
5. Zhu, C.R., Wang, R.S.: Research on Multi-pose Face Image Synthesis from a Single View. *Journal of Electronics and Information Technology* 25(3), 300–305 (2003)
6. Shan, S., Cao, B., Gao, W., Zhao, D.: Extended Fisherface for Face Recognition from a Single Example Image Per Person. In: *IEEE International Symposium on Circuits and Systems, ISCAS 2002*, vol. 2, pp. 81–84. Phoenix-Scottsdale (2002)
7. Zhang, X., Gao, Y., Leung, M.K.: Automatic Texture Synthesis for Face Recognition from Single Views. In: *18th International Conference on Pattern Recognition, ICPR 2006*, vol. 3, pp. 1151–1154 (2006)
8. Yang, C., Shi, S., Li, L., Lv, X.: Face Recognition with Single Training Sample Based on Local Feature Fusion. In: *10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*, pp. 165–169. IEEE (2011)
9. Li, X., Wang, K.J., Be, X.Y.: MW(2D)²PCA Based Face Recognition with Single Training Sample. *Pattern Recognition and Artificial Intelligence* 23(1), 77–83 (2010)
10. Wang, W., Hang, F.F., Li, J.W., Feng, H.L.: Face Description and Recognition by LBP Pyramid. *Journal of Computer-Aided Design & Computer Graphics* 21(1), 94–100 (2009)
11. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: A Robust Local Image Descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1705–1720 (2010)
12. Ahmed, F., Hossain, E., Bari, A.H., Shihavuddin, A.: Compound Local Binary Pattern (CLBP) for Robust Facial Expression Recognition. In: *12th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, pp. 391–395 (2011)
13. Naika, C.L.S., Das, P.K., Nair, S.B.: Asymmetric Region Local Binary Pattern Operator for Person-dependent Facial Expression Recognition. In: *International Conference on Computing, Communication and Applications (ICCCA)*, pp. 1–5. IEEE (2012)
14. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)

A Method of Alteration Information Extracted of Uranium Mine Base on TM Remote Sensing Image

Lin Jiang^{1,2,4}, Yuefei Yi^{3,4}, Shengyu You², Zhibo Wang², and Linlin He²

¹ State Key Laboratory Breeding Base of Nuclear Resources and Environment, East China Institute of Technology, Nanchang, 330013

² Software School, East China Institute of Technology, Nanchang, 330013

³ Department of Computer and Information Science, Hechi University, Yizhou, 546300

⁴ Computer School, Wuhan University, Wuhan 430072

{jlcdf, gxyiyf}@163.com

Abstract. The spectral characteristics are diverse to different mineral substance, therefor the alteration information of remote sensing image is related on the geologic structure features of mineral substance. Extraction of alteration information from remote sensing image, which are contribute to confirm what are consist of geological structure and metallogenic material on diggings area, how to extract the alteration information is all-important to prospecting. This paper proposed a method to extract the alteration information of uranium mine, in our method, the geological structure of uranium and remote data were analyzed in detail, then using band ratio method and principal component analysis(PCA) to TM remote sensing image, which in order to eliminate the negative effect as the alteration information is hybrid by the vegetation index, the final enhancing the alteration information of enhydrite on remote sensing image. Our experiment show, the alteration abnormal on northwestward porphyroclastic lava and in the north of the western in JiangXi Province XiangShan uranium ore field were caused by hydrothermal water mica, which have important implications for uranium prospecting.

Keywords: alteration information, TM remote sensing image, uranium, principal component analysis.

1 Introduction

Since 1990's, remote sensing technology development by leaps and bounds, moreover, image processing, ground-object recognition and Information extraction technologies also rapid development and continuous improvement, which had resulted the extension of application area and intensity. In remote sensing field, the data acquisition from multispectral to hyperspectral, ground-object recognition from secondhand to direct, Information extraction from qualitative to quantitative, which plays a more and more important role in the relate research area such as geological survey, mineral exploration, geological environment evaluation, geological disaster monitoring and some basic geological research.

Wall rock alteration is a geological process which make the new minerals form old disappear by a serial chemical reaction and other appropriate change on hydrothermal between components[1]. Because the extraction of remote sensing abnormal alteration information from remote sensing image, it most impossible that alteration information is a quantifying information about wall rock alteration, which had been applied for prediction of prospecting[2].It is a long time that extraction of alteration information from remote sensing data acquired from Landsat satellite series. As a low resolution of spectrum caused on some pseudo-information while extracted alteration information, which to the disadvantage for prediction of prospecting, in addition, the high resolution of hyperspectral data, but the scan width is narrow, and high price of hyperspectral, which is difficulty for widespread use.

In recent years, Remote sensing alteration information extraction has become a research hotspot. In [1], according to the spectral characteristics of the altered rocks in the ETM image of the study area, the authors enhanced the altered rock information along Wulashan region in InnerMongolia through using the value enhancement, principal component analysis and principal component analysis after enhanced and made extraction and analyses on the alteration information by using the methods of mean plus standard deviation and in combination with the optimistic density separation. Through multi-level selection and assessment, the authors extracted the real anomaly areas and made remote sensing anomaly maps. In [3],the authors based on image enhancement, characteristic vector selection, characteristic vector assemble analysis and anomaly hue distributing rules model dictionary in the studied area is established and alteration information theory is put forward quantitative inverse of anomaly hues is conducted based on alteration information, remarkable prediction results are got in the Xiaoxinancha gold deposit. In [4], based on the relationships of bands of TM image among standard deviation, related coefficient and optimal index factor, the method of optimal index factor combined withspectral signature of alteration information was showed to select the optimal bands composite of remote sensing image, the experiement showed that optimal index factor combined with spectral signature of alteration information was the best way to select the optimal bands composite of multispectral data, structure and lithology of the composite image, which was extracted with the band composite of TM4+TM5+TM7, was clear and different, respectively, band composite TM4+TM5+TM7 was propitious to extract alteration information. In [5], the multi-metal mining are in Lushi, Henan is selected as the studying area, and the ETM+ is the data source, using the method of MPH Masking, Principle component analysis(PCA), Hue and Saturation Transform(HIS) extracts alteration information from the studying area. In [2,6],alteration information was extracted from ASTER and TM image using principal component analysis. In [7,8,9], quantitative extraction method were using to acquire alteration information of deposit, which predicted the distribution of Copper and tungsten on ore deposit.

Above the study, their research focused on extraction of a variety of mineral deposit alteration information, and the image data source from ASTER and TM prevailingly, the extraction method mainly including Quantitative extraction, the main analysis, band ratio method. Regrettably, the extraction of alteration information on

uranium mine there's no involved. Actually, as the extraction of alteration information on uranium mine, we need analysis the uranium ore structure characteristics, and establish the model of quantitative extraction to alteration information, final we should find the prospecting target. In our work, we proposed a method of extraction of alteration information on uranium ore, which are consist of several steps, first, we analyzed the uranium geology structure, interpreted remote sensing image from Landset 7 TM, then extracted the alteration information to uranium ore TM image by using band ratio and principal component analysis, the final we acquire the prediction of minerogenetic region.

2 The Geological Survey of Research on Regional

The Xiangshan uranium ore locates the junction between Chong'ren and Le'an county on Fuzhou city, Jiangxi provence, which is enslaved to the Xiangshan collapse volcanic, and the tectonic position lie at the south side of belt between Yangzi and Chinese ancient plate. There are many achievement about uranium exploration, geological background, volcanic petrology characteristics, ore-controlling factors, metallogenic regularity and metallogenic prediction[10]. From the regional geology of jiangxi province and the blasting vibration and anomaly data of region bouguer gravity, simultaneously, Jiangxi moho depth contour map also show, the Xiangshan ore being located in between edge of the upper mantle uplift and deep tectonic belt. The upwarping region south to Taihe, north to Jingdezhen, west to Xinyu, east to Yingtan, the geological shape like NE. Since the Mesozoic era, the development of continental sedimentary basin and Yanshanian acid in the acid rocks were restricted to the upper mantle uplift belt and deep tectonic belt.

The basement strata of Xiangshan volcanic basin, which control the yield of ore field, is composed of shallow metamorphic rock of Sinian system and under the mountains sandstone of carboniferous China, moreover, The Triassic strata AnYuan group coal measure strata exposed on the east side of volcanic basin. Cross a ravine—Hills the deposit bank deposits located on north of Xiangshan uranium ore, granite porphyry body is a main geological body that controls the yield of uranium ore, the good space of Ore body positioning (group) including internal broken rock altered rock zone, rock mass inside and outside the contact zone, rock mass above the volcanic rocks, and volcanic rock lithology dense fissure zone near the interface, which to control the occurrence of the ore bodies, to ore bodies and tendency to extend for tens to hundreds of meters long, thickness of the general is 1 ~ 5 meters. The average grade of an ore deposit is 0.178% ~ 0.21% [11].

Wall rock alteration of ore deposits has a feature of multi-stage and spatial overlay role. Mine early the albitization development and hydromica alteration, metallogenic epoch the alteration including Hematitization chloritization, hotaru petrochemical and carbonatization, After mineralization mainly carbonatization. The ore type of wall rock alteration which closest, the most obvious is: type uranium—hematite—chlorite, uranium mica type uranium—fluorite—water—type hematite, type uranium—chlorite, uranium - fluorite type. For uranium ore mineral, there are pitchblende, black

titanium, uranium, uranium lambertite, autunite, etc. For metallic mineral, there are pyrite molybdenite galena sphalerite chalcopyrite hematite iron ore red white nickel, etc. Ore of uranium ore is mottled hairline kidney shaped and gelatinous bodies, ore structure are mainly vein veinlet disseminated nests[10,11].

3 The Extraction of Alteration Information

3.1 Data Sources and Pretreatment

Our experiment data is Multispectral come from Landsat-7, the data track number is 121~41, and the time at 4 a.m. July 24,2012, cloudless, data quality is good.

In our image pretreatment experiment, the software is ENVI v4.6.1, we first select polynomial geometric correction method to image rectification, and then overlying the band1 to band7 and display RGB image, beyond that the region of interest(ROI) image was rectified to TM(Clerke, 1986) projection using cubic polynomial transform model. To avoid the Significant degradation of spectral, we resampled the image using nearest neighbor interpolation, pixel size 15 m×15 m(consistent with 8 bands), and control point error X=1.8016m,Y=2.0831m,Total is 2.8021 m, less than 0.5 pixel.

3.2 The Extraction of Alteration Information

Diverse mineral substance contains different spectral characteristics, consequently, research the feature of multi-band spectrum on various kinds alteration. We select uranium ore located on Xiangshan Jiangxi province. According to the previous others research, we know the aqueous mineral alteration information was better reflect the distribution area on Xiangshan uranium ore. So our goal of work is to enhance alteration information of enhydrite from remote sensing image. In our work, we extracted alteration information from TM remote sensing image using band ratio and principal component analysis.

3.2.1 Band Ratio Method

Calculated the ratio of two bands of image, which can weaken background and highlight category or target information, eliminate the impact from terrain, hillshade, cloud, etc. The selection of bands is relate on the feature of multi-bands spectrum from various type of alteration. For instance, the TM3/TM1 would enhance alteration of iron oxide, TM5/TM4 would enhance alteration of ferrous mineral, TM5/TM7 would enhance alteration of carbonatation and chloritization, also TM4/TM3 would enhance vegetation information, on the contrary, TM3/TM4 would restrain it.

The extraction of alteration information on uranium ore focus on enhydrite. According to the feature of enhydrite, the strong peaks appear on band5, and the strongest absorption band appear on band7(near 2.2um), so if enhance band5/band7 would enhance the alteration information of enhydrite, that is the band ratio value is

high if more enhydrite. See fig.1(a), band5/band7 the red layer of upper cretaceous quaternary system and low brightness, while other lithologic differences in brightness value is on the high side but is not big, the band ration is similar to vegetation index image(fig.1(b) band4/band3). So we come to a conclusion that the alteration information of enhydrite and vegetation index had enhanced simultaneously, so the band5/band7 image more highlight than band4/band3 image. In order to extract alteration information ,we using the principal component analysis to do it.

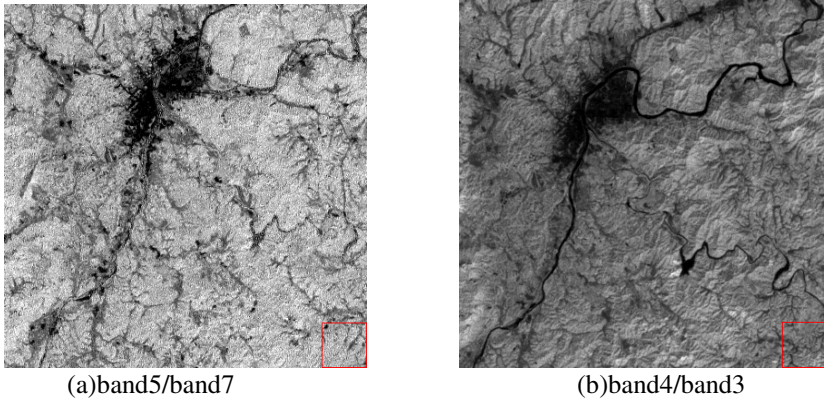


Fig. 1. Comparison of band5/band7 image vs band4/band3 image

3.2.2 Principal Component Analysis

Principal component analysis is a statistics method dimension reduction. By using a orthogonal transformation, it transform the component relevant the original random vector into component is irrelevant the new random vector. On algebraic, it can be view as a transformation from the covariance matrix to diagonal matrix, and on geometry, the original coordinate system was transformed to a new, which pointed p orthogonal direction, then dimension reduction process to multidimensional variable system, this can be transform into a low dimensional variable system from high precision, further the low-dimensional systems is transform into one dimensional system by construct the appropriate value function. The calculation steps following:

(a) Original the standardization of data collection is a p random vector

$$X = (X_1, X_2, \dots, X_p)^T, n \text{ samples } x_i = (x_{i1}, X_{i2}, \dots, X_{ip})^T, i=1,2,\dots,n, n>p,$$

construct sample matrix, and then standardized transformation is as follows:

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, n; j = 1, 2, \dots, p \tag{1}$$

where $\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}.$

(b) Calculate correlation coefficient matrix of standardized matrix Z

$$R = [r_{ij}]_p \quad xp = \frac{Z^T Z}{n-1} \tag{2}$$

where $r_{ij} = \frac{\sum z_{kj} \cdot z_{ki}}{n-1}, i, j, k = 1, 2, \dots, p.$

(c) Solute p eigenvalue of sample correlation matrix from equation $|R - \lambda I_p| = 0$

to determine the principal component, using $\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j} \geq 0.85$ determine m, for

each $\lambda_j, j=1,2,\dots,m$, solve equations set $Rb = \lambda_j b$, we would get unit eigenvector b_j .

(d) Transform target variable into principal component after standardization.

$$U_{ij} = z_i^T b_j^o, j = 1, 2, \dots, m \tag{3}$$

where U_1 is the first principal component, U_2 is the second principal component, ..., U_p is the p^{th} principal component.

(e) Comprehensive assess m principal component Calculate the weighted sum of m principal component, the sum is the final evaluation of estimate, and weight is the rate of contribution of variance each principal component.

3.2.3 PCA Experiment

To avoid the negative influence of vegetation information, we select band5/band7(reflect the alteration information) and band4/band3(reflect vegetation information)to principal component transform, table1 is the eigenvalue after transformation.

Table 1. Statistics of principal component analysis for band5/band7 vs band4/band3

principal component	Eigenvalue	Eigenvector	
		band5/band7	band4/band3
KL1	4959.194983	-0.815480	-0.578785
KL2	167.861670	0.578785	-0.815480

From table1, the correlation coefficient is larger between KL1 and band4/band4,and the correlation coefficient is larger between KL2 and band5/band7, in fact, information of KL2 main derived from band5/band7 ratio image information, namely alteration information of enhydrite. Furthermore, the information contained

KL1 are more larger than KL2, and it shows the first principal component contains most of the variance information of two inputratio image, namely vegetation or non-alteration information; the second principal component enhance the related to poor information of input image, it possible contain alteration information(fig.2(a)).But the noise of KL2 image(fig.2(a)) is large, and many white dots, which results in alteration information is weak and we need do further processing. In fact, select some bands have less geological significance or select PCA to extract enhydrite alteration information is the better strategy.The essence of strategy that extend the spectral contrast of band7 and band5 related on enhydrite to extract alteration information, the precondition is eigenvector load factor must satisfy simultaneously strong negative load of band5 and band7, and opposite the contribution. In our experiment, we select band1,3,5,7 to extract alteration information.

The table2 is eigenvalue after principal component transformation to TM data band1,3,5,7. From table2, we see PC3 has a strong negative load and inverse contribution, which can be use to extract the alteration information of enhydrite. From fig.2(b), the water area is highlight, and the enhydrite region including water mica development zone is more highlight valve.

Table 2. Statistics of principal component analysis for band1、 3、 5、 7

principal component	Eigenvalue	Eigenvector			
		Band1	Band3	Band5	Band7
PC1	2584.564405	0.491043	0.379326	0.660408	0.422906
PC2	172.677103	0.747248	0.277756	-0.484502	-0.36018
PC3	38.360759	-0.236796	0.527603	-0.526206	0.623434
PC4	12.834446	0.380033	-0.70752	-0.228530	0.550228

Apart from band1,3,5,7 PCA, we also do band2,4,5,7 PCA(table3), the result show the fourth principal component KK4 correspond on the extraction condition of enhydrite alteration information. KK4 image(fig.2 (c)) shows the water area or enhydrite development zone is better correspondence with PC3 image, but in KK4 image the light value of water area is the lowest and the light value of enhydrite development zone also lower. In order to obtain better image visual effect, we combing KL2,PC3, KK4 into RGB color image, in our combination image, the green area is water area or enhydrite alteration information zone(fig.2 (d)).

Table 3. Statistics of principal component analysis for band 2、 4、 5、 7

principal component	Eigenvalue	Eigenvector			
		Band2	Band4	Band5	Band7
KK1	3500.658628	0.329985	0.671601	0.563188	0.350544
KK2	207.802688	0.166242	0.638868	-0.52470	-0.537490
KK3	54.881807	0.864040	-0.28100	-0.33030	0.255678
KK4	18.476197	-0.34192	0.248666	-0.54626	0.723085

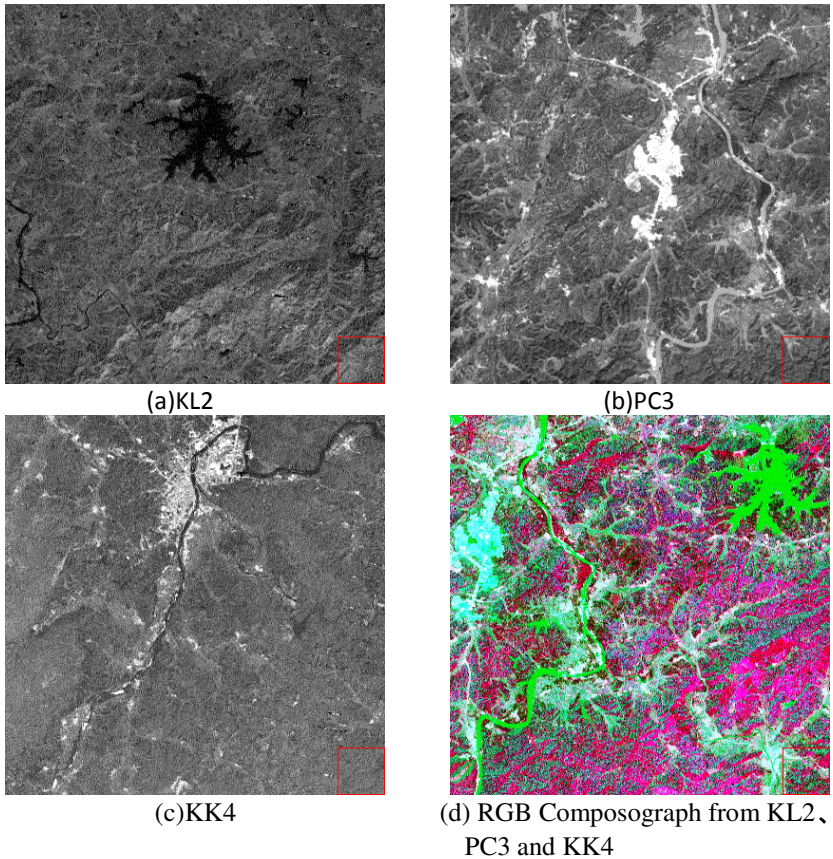


Fig. 2. Comparison of alteration RS anomaly produced using the TM

4 Discussion

The advantage of extraction alteration information from remote sensing image is fast speed, less spend, large processing area, less human disturbance factors, and moreover reflect directly the mineralization alteration information on the surface of the earth. But as multi-spectrum like as TM and ETM+, its band width is large, the spectral difference between alteration rock and background rock is quite weakened because of the domino effect the interval of image band wide spectrum, in addition, it is not neglect to the surface soil and plants. Therefore alteration information existed the background of image as a weak signal. The various type and intensity alteration on TM wide band image is a combined effect, the common enhance and extraction method of information is to distinguish from background image, so it does not find the internal structure and details of information of alteration belt.

5 Conclusion

Xiangshan uranium ore located on subtropical climate zone, and vegetation development better, we acquire a novel visual effect of enhydrite alteration information. Form our experiment find, the alteration information caused by hydrothermal water mica, which would greater contrite to the prospecting.

Acknowledgments. Our work got grant including Jiangxi province education department youth science found (GJJ12397), State Key Laboratory Breeding Base of Nuclear Resources and Environment (East China Institute of Technology)open fund(101114), Guangxi nature science found(2011GXNSFB018074), Guangxi new century teaching reform project(2012JGA198), Guangxi education department scientific research found(2009111x406, 200103YB136) and Hechi college nature science found(2008B-N005).

References

1. Xiang, M., Chen, J., Liu, X.: Study on extraction of remote sensing alteration information in Wulashan. *Jiangsu Geology* 34(1), 28–32 (2010)
2. Hai, D., Da, Z., et al.: The extraction of alteration anomalies from ASTER data based on principal component analysis and fractal model. *Remote Sensing for Land & Resource* 89(2), 75–80 (2012)
3. Lu, F., Xing, L., et al.: Extracting Alteration Information of Remote Sensing Based on Alteration Information Field. *Geology and Prospecting* 42(2), 65–68
4. Wang, P., And Yan, W.: Study on Optimal Bands Composite of TM Image While Extracting Alteration Information. *Journal of Earth Sciences and Environment* 2(6), 173–175 (2010)
5. Liang, Y., And Wang, G.: Alteration From ETM+ Data Rating Based on Fractal Technology. *Remote Sensing Technology And Application* 26(4), 508–511 (2011)
6. Li, M., And Liu, J.: Application of Principal Component Analysis on Alteration Information Extraction By Using TM Remote Sensing Image. *Journal of Natural Science of Helongjiang University* 29(8), 546–550, 555 (2012)
7. Wang, B., Liyan, C.: Ore Prospecting for The Taxkorgan Are of West Kunlun Using Quantitative Extraction of Mineralized Alteration Information. *Geotectonica et Metallogenia* 35(3), 372–377 (2011)
8. Song, W., And Hu, D.: Research On Extraction Of Remote Sensing Alteration Anomalies in JIAMA Copper Deposit In GANGDISE Metallogenic Belt,TIBET. *Journal of Geomechanics* 18(3), 319–330 (2012)
9. Wang, Y., Cui, Y.: The Significance of Abstraction And Study of Remote Sensing Alteration Information of N Part In Ne YUNNAN. *Yunnan Geology* 31(4), 485–488 (2012)
10. Shao, F., Fan, H.: Metallogenic Mechanism And Geochemical Characteristic of Porphyry Type Uranium Deposits, Xiangshan ore field. *Journal of East China Institute of Technology* 34(4), 308–314 (2011)
11. Liu, Z., Liu, Z.: Analysis on Deep Metallogenic Trace And Simulation Experiment in Xiangshan Large-Scale Volcanic Hydrothermal Type Uranium Deposit. *Journal of East China Institute of Technology* 33(3), 201–210 (2010)

A Novel Multi-Scale Local Region Model for Segmenting Image with Intensity Inhomogeneity

Xiao-Feng Wang^{1, 2}, Hai Min^{2, 3}, and Yi-Gang Zhang¹

¹ Key Lab of Network and Intelligent Information Processing, Department of Computer Science and Technology, Hefei University, Hefei Anhui 230601, China

² Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei Anhui 230031, China

³ Department of Automation, University of Science and Technology of China, Hefei, Anhui 230027, China

xfwang@iim.ac.cn, minhai361@gmail.com, yxygz@sina.com

Abstract. In this paper, a novel multi-scale local region (MSLR) model is proposed to segment the image with intensity inhomogeneity. Firstly, the local region is defined in circular shape to capture enough local intensity information. The intensity statistical analysis can be performed in the local circular regions centered in each pixel by using multi-scale average filtering. Secondly, the local term of energy functional is constructed by approximating the difference between intensity statistical information and original image in terms of multi-scale local regions. Thirdly, the regularization term is built to control the smoothness of evolving curve and avoid the over-segmentation phenomenon. Finally, the multi-scale segmentation is performed by minimizing the overall level set energy functional. The experiments on some synthetic and real images with slight or severe intensity inhomogeneity can demonstrate the efficiency and robustness of the proposed MSLR model.

Keywords: image segmentation, intensity inhomogeneity, level set, local region, multi-scale.

1 Introduction

Recently, level set method [1] has become the most popular and successful image segmentation method, in which a deformable curve is represented by the zero level set and driven by evolving the level set functions to approach the objects. Generally, level set method can be divided into edge-based method and region-based method. Edge-based methods [2-3] usually depend on the image gradient to drive the curve evolution and are efficient for segmenting image with sharp gradient. However, they are sensitive to the noise and difficult to detect the weak boundaries. Different from the edge-based methods, the region-based methods [4-6] utilize global region information to stabilize their responses to local variations and guide the evolving curve to approach the objects. Among them, Chan-Vese (CV) model [4] is the most representative and popular one, which approximates a given image with constant functions

inside each region. However, the constant function can not precisely describe the intensity distribution of region with intensity inhomogeneity. Thus, it often fails to segment the images with intensity inhomogeneity.

To improve the performance of CV model, some local region-based models [7-9] have been proposed, e.g. the local binary fitting (LBF) model [7], the local image fitting (LIF) model [8] and the local Chan-Vese (LCV) model [9], etc. These models generally adopt the assumption that the images with intensity inhomogeneity are intensity homogeneous in smaller local region. Thus, they can segment the images with slight intensity inhomogeneity by approximating the given image in terms of local region rather than global region. In practical implementation, they usually use certain statistical function with fixed scale to examine the local region centered in each pixel. Due to the non-linear inhomogeneity, fixing scale for all local regions may result in the failed segmentation for image with severe intensity inhomogeneity. To improve the segmentation performance, the multiple scales should be used.

In this paper, we proposed a so-called multi-scale local region (MSLR) model to segment image with intensity inhomogeneity by introducing the multi-scales local statistical analysis into the local region method. Firstly, the local region is defined in circular shape to capture enough local intensity information. Then, the intensity statistical analysis shall be performed in the local circular regions centered in each pixel by using multi-scale average filtering. Thus, the multi-scale intensity statistical information is extracted. The local term of energy functional can be constructed by approximating the difference between multi-scale intensity statistical information and original image. To further control the smoothness of the evolving curve and avoid the over-segmentation phenomenon, the regularization term is included in the energy functional. Finally, the multi-scale segmentation is performed by minimizing the new formed level set energy functional.

The rest of this paper is organized as follows: In Section 2, we briefly review the CV model and LCV model. Our multi-scale local region (MSLR) model is presented in Section 3. In Section 4, the proposed model is validated by some experiments on synthetic and real images with slight or severe intensity inhomogeneity. Finally, some conclusive remarks are included in Section 5.

2 Background Knowledge

2.1 Chan-Vese Model

The Chan-Vese (CV) model [4] is to minimize the following energy functional in terms of the level set function $\phi(x, y)$.

$$\begin{aligned}
 E_{\epsilon}^{CV}(c_1, c_2, \phi) = & \mu \cdot \int_{\Omega} \delta_{\epsilon}(\phi(x, y)) |\nabla \phi(x, y)| dx dy \\
 & + \lambda_1 \cdot \int_{\Omega} |I(x, y) - c_1|^2 H_{\epsilon}(\phi(x, y)) dx dy + \lambda_2 \cdot \int_{\Omega} |I(x, y) - c_2|^2 (1 - H_{\epsilon}(\phi(x, y))) dx dy,
 \end{aligned} \tag{1}$$

where λ_1, λ_2 and μ are positive constants. c_1 and c_2 are intensity averages of given image I inside and outside evolving curve, respectively. $H_\epsilon(z)$ and $\delta_\epsilon(z)$ are the regularized approximation of Heaviside function $H(z)$ and Dirac delta function $\delta(z)$, respectively.

The main advantages of CV model are it can detect objects whose boundaries are not necessarily defined by gradient and can efficiently segment the noisy images. However, CV model is primitively designed for segmenting images with intensity homogeneity since it assumes that the intensities always maintain constant in each region. Thus, CV model often leads to poor segmentation results for images with intensity inhomogeneity due to wrong movement of evolving curves guided by global image information.

2.2 Local Chan-Vese Model

The Local Chan-Vese (LCV) model is proposed by us in [9], which incorporates the statistical information extracted from local square region into the energy functional. The overall energy functional in terms of level set function ϕ is written as follows:

$$\begin{aligned}
& E_\epsilon^{LCV}(c_1, c_2, d_1, d_2, \phi) \\
&= \int_{\Omega} (\alpha \cdot |I(x, y) - c_1|^2 + \beta \cdot |g_k * I(x, y) - I(x, y) - d_1|^2) H_\epsilon(\phi(x, y)) dx dy \\
&\quad + \int_{\Omega} (\alpha \cdot |I(x, y) - c_2|^2 + \beta \cdot |g_k * I(x, y) - I(x, y) - d_2|^2) (1 - H_\epsilon(\phi(x, y))) dx dy \quad (2) \\
&\quad + (\mu \cdot \int_{\Omega} \delta_\epsilon(\phi(x, y)) |\nabla \phi(x, y)| dx dy + \int_{\Omega} \frac{1}{2} (|\nabla \phi(x, y) - 1|)^2),
\end{aligned}$$

where g_k is an averaging convolution operator with $k \times k$ size window.

The successful segmentation of LCV model for image with slight intensity inhomogeneity relies on the assumption that smaller local regions are more likely to have approximately homogeneous intensity and the intensity of the object is statistically different from the background [9]. However, the scale k of LCV model is still fixed for all local square region of whole image. Although the assumption is still tenable for images with severe intensity inhomogeneity, the most ideal scale for all local regions can hardly be found to support the assumption. Thus, the dissatisfying segmentation result may produced by LCV model for images with severe intensity inhomogeneity (as shown in the Fig.3 and Fig.4).

3 Multi-Scale Local Region Model

In this section, we shall present and discuss the details of our multi-scale local region (MSLR) model. The overall energy functional is composed of local term E^L and regularization term E^R .

3.1 Local Term

The local term is the key to improve the segmentation performance for images with severe intensity inhomogeneity in our model. Intensity inhomogeneity is known as the systematic intensity change on both the foreground and background regions. It manifests itself as a smooth spatially varying function that changes image intensities which would be originally constant for the same tissue type regardless of its position in the image [10]. Traditional region-based level set methods are highly sensitive to the spurious variations of intensities since they assume that the intensities in each region are constant. Actually, the spectrum of intensity inhomogeneity in frequency domain is concentrated in the low-frequency area [9]. The intensity inhomogeneity effect shall mainly change the intensity distribution of non-contour pixels, whereas for that of contour pixels, the change is relatively small. So, it is significant to statistically analyze each pixel with respect to its local neighborhood. In other words, the statistical analysis should be performed in local regions centered in each pixel.

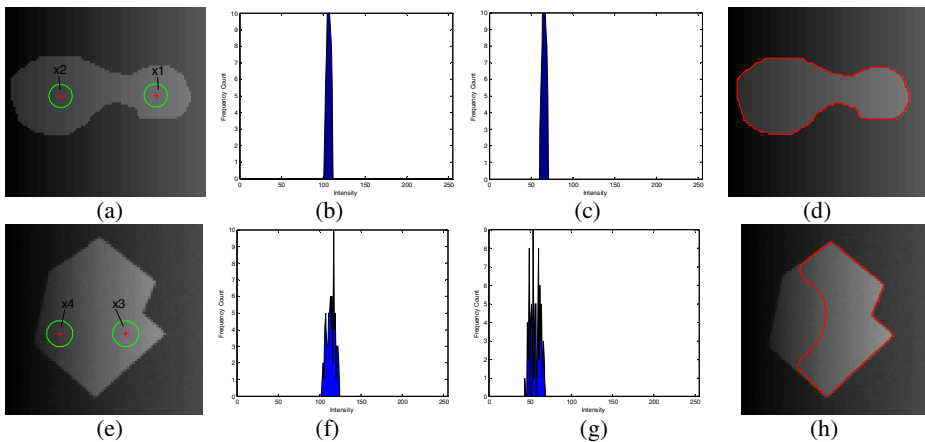


Fig. 1. Segmentation for images with slight and severe intensity inhomogeneity by using LCV model. (a) Image with slight intensity inhomogeneity. (b) The intensity distribution of local region centered in x1. (c) The intensity distribution of local region centered in x2. (d) Successful segmentation result. (e) Image with severe intensity inhomogeneity. (f) The intensity distribution of local region centered in x3. (g) The intensity distribution of local region centered in x4. (h) Failed segmentation result.

As mentioned above, fixing scale for all local regions is improper for segmenting the image with severe intensity inhomogeneity. Here, we shall provide an experiment to verify the conclusion. In Fig.1, we demonstrate the segmentation for images with slight and severe intensity inhomogeneity by using LCV model. Fig.1 (a) shows the well-known test image with slight intensity inhomogeneity. We labeled two pixels x1 and x2 on the image where red crosses denote the pixels and green circles denote their local regions with scale of 8. It can be seen from the intensity distribution histograms in Fig.1 (b) and (c) that two local regions centered in x1 and x2 both have the approximately homogeneous intensity. Thus, Fig.1 (a) can be easily segmented by LCV

model (as shown in Fig.1 (d)). Fig.1 (e) shows an image with severe intensity inhomogeneity. We also labeled two pixels x_3 and x_4 on the image. Fig.1 (e) shows that local region centered in x_3 has a little inhomogeneity which is different from Fig.1 (b) and (c). Fig.1 (f) shows that local region centered in x_4 has much more inhomogeneity. Fig.1 (h) shows that LCV model fails to segment the image especially in left region where severe intensity inhomogeneity occurs. To segment this image, the scales should be different for local regions at the left and right side in object.

Now, we shall present our local term which performs the multi-scale intensity statistical analysis in local regions. The local region in our model is defined in circular shape which is different from the square shape used in other local region-based methods. The purpose is to capture more local statistical intensity information since the non-uniform illumination, which may cause the intensity inhomogeneity phenomenon, is usually circular scattered in the shape. So, the intensity statistical analysis shall be performed in the local circular regions with multiple scales.

Generally speaking, the spectrum of intensity inhomogeneity in frequency domain is concentrated in the low-frequency area. That is to say, intensity inhomogeneity can be regarded as a low-frequency artifact and further be separated from the high-frequency object contour by certain low-pass filtering. Hence, we consider choosing the low-pass filtering to perform the multi-scale intensity statistical analysis. Among the low-pass filtering methods, the most fast one is the average filtering method. Since that the object contour pixels should be brighter than the average if they are brighter than the background, the average filtering can also work effectively for analyzing the local circular region in our model.

Here, we firstly construct the multi-scale average filtering to analyze the local circular region centered in each pixel of the given image I .

$$AF_i(x, y) = \frac{1}{n} \sum_{(x', y') \in R_i} I(x', y') \quad R_i : \sqrt{(x' - x)^2 + (y' - y)^2} \leq i, \quad (3)$$

where $AF_i(x, y)$ is the average filter whose scale is the subscript i . n denotes the number of pixels belonging to the local circular region R_i which is centered in (x, y) .

Then, the mean of the multi-scale average filtering $M_k(x, y)$ can be computed and regarded as the multi-scale intensity statistical information.

$$M_k(x, y) = \frac{1}{k} \sum_{i=1}^k AF_i(x, y), \quad (4)$$

where k denotes the number of scales. To capture enough statistical information, k is usually set to a larger value. In this paper, we fix $k = 32$ for all experiments.

By subtracting the original image $I(x, y)$ from $M_k(x, y)$, the contrast between intensities of object contour and background can be significantly increased. It means that the separation of object contour and background may be performed even in the image with severe intensity inhomogeneity. Here, the local term of energy functional is constructed by approximating the difference $(M_k(x, y) - I(x, y))$ as follows:

$$\begin{aligned}
 E^L(m_1, m_2, C, k) = & \int_{\text{inside}(C)} |M_k(x, y) - I(x, y) - m_1|^2 dx dy \\
 & + \int_{\text{outside}(C)} |M_k(x, y) - I(x, y) - m_2|^2 dx dy,
 \end{aligned}
 \tag{5}$$

where m_1 and m_2 are the intensity averages of the difference $(M_k(x, y) - I(x, y))$ inside C and outside C , respectively.

The local term (5) can be reformulated in terms of the level set function $\phi(x, y)$ as:

$$\begin{aligned}
 E^L(m_1, m_2, \phi, k) = & \int_{\Omega} |M_k(x, y) - I(x, y) - m_1|^2 H(\phi(x, y)) dx dy \\
 & + \int_{\Omega} |M_k(x, y) - I(x, y) - m_2|^2 (1 - H(\phi(x, y))) dx dy.
 \end{aligned}
 \tag{6}$$

3.2 Regularization Term

To control the smoothness of evolving curve C , the length penalty term $L(C)$ should be included in the regularization term. The use of $L(C)$ implies that the evolving curve minimizing the energy functional should be as short as possible. By replacing the curve C with the level set function $\phi(x, y)$, $L(C)$ can be reformulated as:

$$L(\phi = 0) = \int_{\Omega} \delta(\phi(x, y)) |\nabla \phi(x, y)| dx dy.
 \tag{7}$$

It should be noted that the over-segmentation phenomenon often exists in the segmentation for image with intensity inhomogeneity. Evolving curve may split and approach some small and isolated regions which should belong to the background. To avoid this phenomenon, the parameter μ should be added to control the penalization effect of $L(\phi = 0)$. It is particularly important in our model for segmenting the image with severe intensity inhomogeneity. Generally speaking, small objects will be detected if μ is small and larger objects are detected if μ is larger. The parameter μ should be elaborately selected according to the segmentation purpose in practical implementation. In addition, to naturally keep the level set function as a signed distance function in the evolution, the penalty term $P(\phi)$ proposed in [7] is also included in the regularization term of our model.

Therefore, the regularization term E^R is written as the following:

$$E^R(\phi) = \mu \cdot \int_{\Omega} \delta(\phi(x, y)) |\nabla \phi(x, y)| dx dy + \int_{\Omega} \frac{1}{2} (|\nabla \phi(x, y)| - 1)^2.
 \tag{8}$$

3.3 Level Set Formulation

Finally, the overall energy functional of MSLR model can be described as follows:

$$\begin{aligned}
 E_{\varepsilon}^{MSLR}(m_1, m_2, \phi, k) &= E^L(m_1, m_2, \phi, k) + E^R(\phi) \\
 &= \int_{\Omega} |M_k(x, y) - I(x, y) - m_1|^2 H_{\varepsilon}(\phi(x, y)) dx dy \\
 &\quad + \int_{\Omega} |M_k(x, y) - I(x, y) - m_2|^2 (1 - H_{\varepsilon}(\phi(x, y))) dx dy \\
 &\quad + \mu \cdot \int_{\Omega} \delta_{\varepsilon}(\phi(x, y)) |\nabla \phi(x, y)| dx dy + \int_{\Omega} \frac{1}{2} (|\nabla \phi(x, y) - 1|)^2,
 \end{aligned} \tag{9}$$

where $H_{\varepsilon}(z)$ and $\delta_{\varepsilon}(z)$ can be described as follows:

$$H_{\varepsilon}(z) = \frac{1}{2} \left| 1 + \frac{2}{\pi} \arctan \left| \frac{z}{\varepsilon} \right| \right|, \quad \delta_{\varepsilon}(z) = \frac{1}{\pi} \cdot \frac{\varepsilon}{\varepsilon^2 + z^2}. \tag{10}$$

By calculus of variations, the constant functions $m_1(\phi)$ and $m_2(\phi)$ that minimize $E^{MSLR}(m_1, m_2, \phi, k)$ for a fixed function ϕ are given by:

$$m_1(\phi) = \frac{\int_{\Omega} (M_k(x, y) - I(x, y)) H_{\varepsilon}(\phi(x, y)) dx dy}{\int_{\Omega} H_{\varepsilon}(\phi(x, y)) dx dy}. \tag{11}$$

$$m_2(\phi) = \frac{\int_{\Omega} (M_k(x, y) - I(x, y)) (1 - H_{\varepsilon}(\phi(x, y))) dx dy}{\int_{\Omega} (1 - H_{\varepsilon}(\phi(x, y))) dx dy}. \tag{12}$$

Keeping m_1 and m_2 fixed and minimizing the overall energy function E^{MSLR} with respect to ϕ , we can deduce the associated Euler-Lagrange equation for ϕ .

$$\begin{aligned}
 \frac{\partial \phi}{\partial t} &= \delta_{\varepsilon}(\phi) [(M_k(x, y) - I(x, y) - m_2) - (M_k(x, y) - I(x, y) - m_1)^2] \\
 &\quad + [\mu \cdot \delta_{\varepsilon}(\phi) \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) + (\nabla^2 \phi - \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right))].
 \end{aligned} \tag{13}$$

To solve the above partial differential equation, the finite difference method is adopted in this paper.

4 Experimental Results

In this section, we shall present the experimental results of our MSLR model on some synthetic and real images with slight and severe intensity inhomogeneity. The proposed model was implemented by Matlab 7 on a computer with Intel Core 2 Duo

2.2GHz CPU, 2G RAM, and Windows 7 operating system. The processing time referred in this section starts after manually setting the initial contour. Here, we used the same parameters of the time-step $\Delta t = 0.1$, the grid spacing $h = 1$, $\varepsilon = 1$ (for $H_\varepsilon(z)$ and $\delta_\varepsilon(z)$) and the number of scales $k = 32$ for all the experiments. The value of μ controlling the length penalization effect should be adjusted according to practical situation.

Firstly, we used the synthetic images with slight intensity inhomogeneity to validate the performance of the proposed MSLR model. The first column of Fig.2 shows two synthetic images with slight intensity inhomogeneity. It can be seen that the intensity increases gradually from left to right in the first image and increases gradually from top to bottom in the second image. The initial contours were placed at the intersectional region of high intensity area and low intensity area, as shown in the second column of Fig.2. The final segmentation results of the MSLR model are demonstrated in the third column of Fig.2. The evolving curve successfully surrounded the object at the 4th iteration and the 12th iteration, respectively.

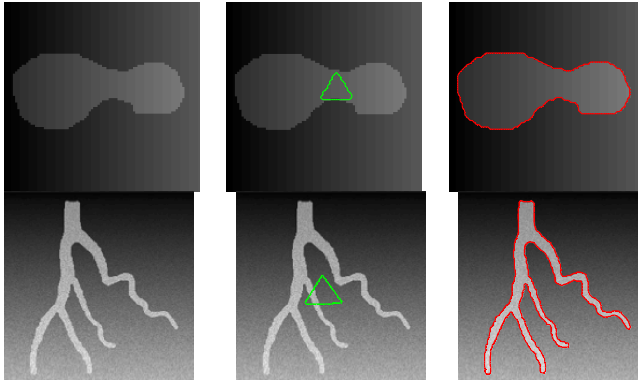


Fig. 2. Segmentation for synthetic images with slight intensity inhomogeneity by using MSLR model. The first column: Original images. The second column: Initial contours. The third column: Final segmentation result of the MSLR model at the 4th iteration and the 12th iteration. Size=88×85, 176×177. $\mu = 0.001$ and 0.01. Processing times=1.24s and 6.65s.

Next, we shall demonstrate the ability of the proposed MSLR model to segmenting synthetic images with severe intensity inhomogeneity. To make a comparison with traditional local region-based methods, the LCV model was adopted for segmenting the images. Fig.3 shows the segmentation for two synthetic images with severe intensity inhomogeneity using both LCV model and MSLR model. The first column shows that the initial contours were placed at the bottom left corner where LCV model can achieve good performance. The second column of Fig.3 shows that LCV model failed to segment both two images. The evolving curves could not cross the intersection of high intensity region and low intensity region. The third column of Fig.3 shows that MSLR model successfully segmented two images at the 43rd iteration and the 44th iteration, respectively.

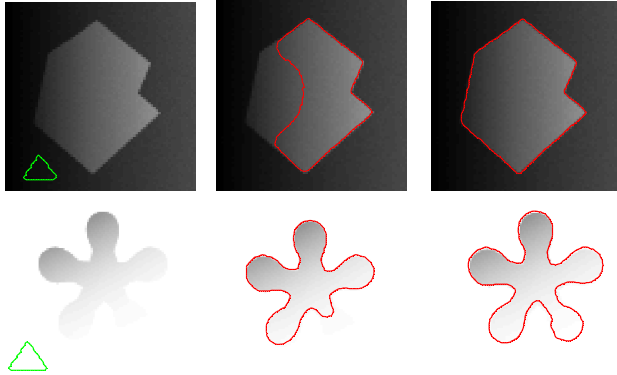


Fig. 3. Comparisons of LCV model and MSLR model on segmenting synthetic images with severe intensity inhomogeneity. The first column: Initial contours. The second column: Final segmentation results using LCV model. The third column: Final segmentation results using the proposed MSLR model at the 43rd iteration and 44th iteration. Size= 76×76 , 100×100 . $\mu = 0.001$ and 0.0001. Processing time=1.24s and 6.65s.

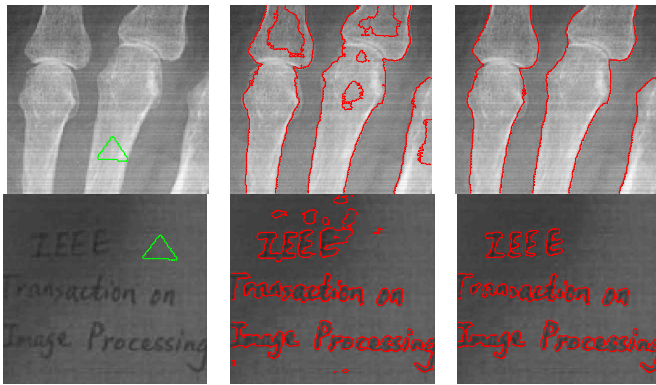


Fig. 4. Comparisons of LCV model and MSLR model on segmenting real images with slight and severe intensity inhomogeneity. The first column: Initial contours. The second column: Final segmentation results using LCV model. The third column: Final segmentation results using the proposed MSLR model at the 50th iteration and the 20th iteration. Size= 126×117 , 145×135 . $\mu = 0.02$ and 0.001. Processing time=8.39s and 4.17s.

In Fig.4, we illustrate the segmentation of LCV model and MSLR model for two real images. The first image is a medical image with slight intensity inhomogeneity and the second image is a hand-writing image with severe intensity inhomogeneity. The first column of Fig.4 shows that the initial contours were placed on the positions where LCV model can achieve the best performance. It can be seen from the second column of Fig.4 that LCV model failed to segment both two images. The over-segmentation phenomenon occurred while segmenting the first image. For the second image, the last hand-writing letter E in abbreviative word IEEE was not segmented. Besides, the over-segmentation phenomenon also occurred. The third column of Fig.4

shows the final segmentation results using MSLR model at the 50th iteration and the 20th iteration. It can be seen that both two real images were successfully segmented.

5 Conclusions

In this paper, we proposed a novel multi-scale local region (MSLR) model for segmenting the image with intensity inhomogeneity. Firstly, the intensity statistical analysis is performed in local circular regions centered in each pixel by multi-scale average filtering. Then, the local term is constructed to approximate the difference between extracted intensity statistical information and original image in terms of multi-scale local regions. Finally, the multi-scale segmentation is performed by minimizing the overall level set energy functional. Experiments on some synthetic and real images can demonstrate that MSLR model is efficient and robust for segmenting image with slight or severe intensity inhomogeneity.

Acknowledgements. This work was supported by the grant of the National Natural Science Foundation of China, No. 61005010, the special grant of China Postdoctoral Science Foundation, No. 2012T50582, the grant of Anhui Provincial Natural Science Foundation, No. 1308085MF84, the grant of China Postdoctoral Science Foundation, No. 20100480708, the grant of the Key Scientific Research Foundation of Education Department of Anhui Province, No. KJ2010A289, the grant of Scientific Research Foundation for Talents of Hefei University, No. 11RC05.

References

1. Osher, S., Sethian, J.A.: Fronts Propagating with Curvature-Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *J. Comput. Phys.* 79(1), 12–49 (1988)
2. Caselles, V., Catte, F., Coll, T., Dibos, F.: A Geometric Model for Active Contours in Image Processing. *Numer. Math.* 66(1), 1–31 (1993)
3. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. *Int. J. Comput. Vision* 22(1), 61–79 (1997)
4. Chan, T.F., Vese, L.A.: Active Contours without Edges. *IEEE Trans. Image Processing* 10(2), 266–277 (2001)
5. Paragios, N., Deriche, R.: Geodesic Active Regions and Level Set Methods for Supervised Texture Segmentation. *Int. J. Comput. Vision* 46(4), 223–247 (2002)
6. Gao, S., Bui, T.D.: Image Segmentation and Selective Smoothing by Using Mumford–Shah Model. *IEEE Trans. Image Processing* 14(10), 1537–1549 (2005)
7. Li, C., Kao, C., Gore, J.C., Ding, Z.: Minimization of Region-Scalable Fitting Energy for Image Segmentation. *IEEE Trans. Image Processing* 17, 1940–1949 (2008)
8. Zhang, K.H., Song, H.H., Zhang, L.: Active Contours Driven by Local Image Fitting Energy. *Pattern Recognition* 43(4), 1199–1206 (2010)
9. Wang, X.F., Huang, D.S., Xu, H.: An Efficient Local Chan-Vese Model for Image Segmentation. *Pattern Recognition* 43(3), 603–618 (2010)
10. Vovk, U., Pernuš, F., Likar, B.: A Review of Methods for Correction of Intensity Inhomogeneity in MRI. *IEEE Trans. Med. Imag.* 26(3), 405–421 (2007)

Fault Tolerant Control for Robot Manipulators Using Neural Network and Second-Order Sliding Mode Observer

Mien Van¹ and Hee-Jun Kang^{2,*}

¹Graduate School of Electrical Engineering, University of Ulsan, 680-749, South Korea
varmien1@gmail.com

²School of Electrical Engineering, University of Ulsan, 680-749, South Korea
hjkang@ulsan.ac.kr

Abstract. This paper investigates an algorithm for fault tolerant control of uncertain robot manipulator with only joint position measurement using neural network and second-order sliding mode observer. First, a neural network (NN) observer is designed to estimate the modeling uncertainties. Based on the obtained uncertainty estimation, a second-order sliding mode observer is then designed for two purposes: 1) Providing the velocity estimation, 2) providing the fault information that is used for fault detection, isolation and identification. Finally, a fault tolerant control scheme is proposed for compensating the effect of uncertainties and faults based on the fault estimation information. Computer simulation results on a PUMA560 industrial robot are shown to verify the effectiveness of the proposed strategy.

Keywords: Fault diagnosis, Neural network, Second-order sliding mode observer, Robot dynamics.

1 Introduction

Various fault diagnosis approaches for nonlinear systems as well as robotic system have been studied during the last three decades. The model-based analytical redundancy based fault detection and isolation have been widely investigated in the literature [1]. By using neural network (NN) learning, a robust fault diagnosis schemes have been proposed [2-3]. Other observer based technique which is so-called sliding mode (SM) observer have been developed, in which the equivalent output injection (EOI) signal is utilized to detect and reconstruct the presence of the uncertainties and faults [4]. In some applications of sliding mode, chattering is the major drawbacks of the SM approach in the practical realization. The most widely used in practical applications to eliminate chattering are using higher-order sliding mode such as second-order sliding mode (SOSM) [5-6, 10].

After a fault has been detected and isolated, in some applications of robotic system it is required that the fault must be self-corrected, to guarantee that the robot can

* Corresponding author.

continue working. It is referred as fault tolerant control (FTC). In general, FTC can be divided into two main approaches [7]: 1) passively, one controller is used for both normal case and fault case without the need to detect the presence of the fault [8]. This approach however, requires the partial knowledge of possible system faults. It is limited in real application; 2) actively, FTC is designed based on fault diagnosis information [2-3, 9-10]. In the actively FTC, fault diagnosis (FD) is the first step to provide the fault information. Based on the obtained fault information, a FTC is then designed to compensate for the effect of the faults in the system by online controller reconfiguration. With the correct fault information, performance of an active FTC system is more effective than that of a passive FTC system and hence is more desirable for practical applications. For robotic system, a number of FTC approaches have been proposed based on active strategy [2-3, 9-10]. However, these approaches did not compensate for the modeling uncertainty. Hence, the tracking performance is decreased. In addition, these schemes were designed based on the assumption that the joint velocity measurements are available. It is limited in real application. To overcome this obstacle, exact velocity observer has proposed based on super-twisting SOSM observer [6, 13].

In light of the remarkable benefits, in this paper, an active FTC scheme is proposed for uncertain robot manipulators using only position measurement. In the proposed scheme, both uncertainty and fault are considered and compensated. The FTC scheme is constructed based on the fault estimation which is obtained from a fault diagnosis observer scheme. The main contribution of this paper is as follows: 1) a second-order sliding mode observer is designed to obtain both velocity and fault estimation, 2) a neural network observer is used to estimate the uncertainties, 3) an active FTC scheme is designed based on the velocity estimation, uncertainty estimation and fault estimation for compensating the effects of the uncertainties and faults. Simulation results using a PUMA560 robot arm are used to verify the effectiveness of the proposed algorithm.

2 Problem Statement

Consider the robot dynamics is described by

$$\ddot{q} = M^{-1}(q)[\tau - V_m(q, \dot{q})\dot{q} - G(q) - \Delta'(q, \dot{q}, t)] + \beta(t - T_f)\phi(q, \dot{q}, \tau) \tag{1}$$

where $q \in \mathfrak{R}^n$ is the state vector, $\tau \in \mathfrak{R}^n$ is the torque produced by the actuators, $M(q) \in \mathfrak{R}^{n \times n}$ is the inertia matrix. $V_m(q, \dot{q}) \in \mathfrak{R}^n$ is the Coriolis and centripetal torque, $G(q) \in \mathfrak{R}^n$ is the term of gravitational torque, $\Delta'(q, \dot{q}, t) \in \mathfrak{R}^n$ is an uncertainty term. The term $\phi(q, \dot{q}, \tau) \in \mathfrak{R}^n$ is a vector which represents the changes in the system dynamics due to the occurrence of a fault, $\beta(t - T_f) = \text{diag}\{\beta_1(t - T_f), \beta_2(t - T_f), \dots, \beta_n(t - T_f)\} \in \mathfrak{R}^n$ represents the time profile of the faults, and T_f is the time of occurrence of the faults, that is

$$\beta_i(t-T_f) = \begin{cases} 0 & \text{if } t < T_f \\ 1 - e^{-\varphi_i(t-T_f)} & \text{if } t \geq T_f \end{cases} \quad (2)$$

where $\varphi_i > 0$ represent the unknown fault evolution rate. Small φ_i values characterize slowly developing forms, also called incipient fault. For very large values of φ_i , the profile of β_i approaches a step function that models abrupt faults. When $\varphi_i \rightarrow \infty$, the β_i become a step function so that incipient fault becomes an abrupt faults.

To simply the subsequent design and analysis, Eq. (1) can be rewritten as

$$\ddot{q} = M^{-1}(q)[\tau - H(q, \dot{q})] + \Delta(q, \dot{q}, t) + \beta(t-T_f)\phi(q, \dot{q}, \tau) \quad (3)$$

where $H(q, \dot{q}) = V(q, \dot{q}) + G(q)$ and $\Delta(q, \dot{q}, t) = -M^{-1}(q)\Delta'(q, \dot{q}, t)$.

In this paper, we investigate a FTC scheme to handle the effects of faults of uncertainty robot manipulator rely on the following assumptions.

Assumption 1: The states of the robot system are bounded for all time.

Assumption 2: The modeling uncertainty is bounded

$$|\Delta(q, \dot{q}, t)| < \bar{\Delta} \quad (4)$$

where $\bar{\Delta}$ is a known constant.

Assumption 3: The unknown fault function is bounded

$$|\phi(q, \dot{q}, t)| < \bar{\phi} \quad (5)$$

where $\bar{\phi}$ is a known constant.

By introducing $x_1 = q \in \mathfrak{R}^n$, $x_2 = \dot{q} \in \mathfrak{R}^n$, the robot dynamics expressed in eq. (3) can be written in state space form as:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= f(x_1, x_2, u) + \Delta(x_1, x_2, t) + \beta(t-T_f)\phi(x_1, x_2, u) \\ y &= x_1 \end{aligned} \quad (6)$$

where $u = \tau$, $f(x_1, x_2, u) = M^{-1}(q)[\tau - H(q, \dot{q})]$ denotes the nominal of robot dynamics.

3 Design of Uncertainty Observer

The uncertainty $\Delta(x_1, x_2, t)$ can be described by a neural network approximation as:

$$\Delta(x_1, x_2, t) = W\sigma(V\bar{x}(t)) + \delta \quad (7)$$

where W, V are the ideal weights of three layer neural network, $\bar{x}(t) = [x_1, x_2, u]$ is the neural network input, σ is sigmoid activation function and δ is approximation error.

In order to estimate the uncertainties, a neural network based observer is designed as

$$\dot{\omega} = A(\omega - x_2) + f(x_1, \omega, u) + U(t) \tag{8}$$

where $A = \text{diag}\{-k_1, -k_2, \dots, -k_n\}$ is stable matrix with $k_i > 0$, ω denotes the state of the observer. $U(t)$ is the observer input which is used to estimate the uncertainties and it can be selected by

$$U(t) = \hat{W} \sigma(\hat{V} \bar{x}(t)) \tag{9}$$

where \hat{W} and \hat{V} are the estimation of the ideal weights W and V , respectively.

By chosen a cost function as $J = \frac{1}{2}(\tilde{\omega})^T(\tilde{\omega})$, $\tilde{\omega} = x_2 - \omega$, the update law based on the back-propagation plus an e-modification terms is designed as [12]:

$$\dot{\hat{W}} = -\rho_1 \frac{\partial J}{\partial \hat{W}} - \rho_2 \|\tilde{\omega}\| \hat{W} \tag{10}$$

$$\dot{\hat{V}} = -\rho_3 \frac{\partial J}{\partial \hat{V}} - \rho_4 \|\tilde{\omega}\| \hat{V} \tag{11}$$

where $\rho_1, \rho_3 > 0$ are the learning rates and $\rho_2, \rho_4 > 0$ are small positive numbers. Substituting Eq. (6) into Eq. (8) when the robot works in normal operation mode, the observer error is defined as:

$$\dot{\tilde{\omega}} = A\tilde{\omega} + f(x_1, x_2, u) - f(x_1, \omega, u) + \Delta(x_1, x_2, t) - U(t) \tag{12}$$

By similar way with Ref. [10], it is easy to demonstrate that the estimation error ($\tilde{\omega}$) and the resulting parameter errors ($\tilde{W} = W - \hat{W}$, $\tilde{V} = V - \hat{V}$) are stability and being bounded under the observer scheme expressed in Eq. (8).

After the NN weights vector \hat{W}, \hat{V} approach to the optimal vectors W, V under the update laws expressed in Eqs. (10) and (11) in normally working condition of robot system. At this time, the NN weights are fixed and noted as W_0, V_0 . Up to now, the uncertainty can be expressed as:

$$\Delta(x_1, x_2, t) = W^0 \sigma(V^0 \bar{x}(t)) + \varepsilon \tag{13}$$

and the uncertainty estimation expressed in eq. (7) can be replaced by obtained NN:

$$\hat{\Delta}(x_1, x_2, t) = W^0 \sigma(V^0 \bar{x}(t)) \tag{14}$$

where $\hat{\Delta}(x_1, x_2, t)$ denotes the estimation of the uncertainty $\Delta(x_1, x_2, t)$.

4 Velocity and Fault Diagnosis Observer Scheme

Based on [6], the second-order sliding mode based state observer can be described by

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + \lambda |\tilde{x}_1|^{1/2} \text{sign}(\tilde{x}_1) \\ \dot{\hat{x}}_2 &= f(x_1, \hat{x}_2, u) + \hat{\Delta}(x_1, x_2, t) + \alpha \text{sign}(\tilde{x}_1) \end{aligned} \tag{15}$$

where \hat{x}_1 and \hat{x}_2 are states estimation of x_1 and x_2 with initial condition $\hat{x}_1 = x_1$, $\hat{x}_2 = 0$. $\tilde{x}_1 = x_1 - \hat{x}_1$ is the state estimation error.

Substituting Eq. (6) into Eq. (15), we obtain the states estimation error:

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{x}_2 - \lambda |\tilde{x}_1|^{1/2} \text{sign}(\tilde{x}_1) \\ \dot{\tilde{x}}_2 &= d(x_1, x_2, \hat{x}_2, u) + \varepsilon + \phi(x_1, x_2, u) - \alpha \text{sign}(\tilde{x}_1) \end{aligned} \tag{16}$$

where $d(x_1, x_2, \hat{x}_2, u) = f(x_1, x_2, u) - f(x_1, \hat{x}_2, u)$, $\varepsilon = \Delta(x_1, x_2, t) - \hat{\Delta}(x_1, x_2, t)$ is the uncertainty estimation error, it is bounded by $|\varepsilon| < \bar{\varepsilon}$, $\bar{\varepsilon}$ is a known constant. If we defined $F(x_1, x_2, \hat{x}_2, u, t) = d(x_1, x_2, \hat{x}_2, u) + \varepsilon + \phi(x_1, x_2, u)$, based on the assumptions 1-3, there exist a constant f^+ such that:

$$|F(x_1, x_2, \hat{x}_2, u, t)| < f^+ \tag{17}$$

Based on the Lyapunov approach in Ref. [13], Theorem 1 gives the convergence of the estimation error to zero in finite time:

Theorem 1: Suppose that condition (17) holds for a system as in Eq. (16); if the sliding mode gains of the observer scheme in Eq. (16) are chosen as

$$\begin{aligned} \lambda &> 0 \\ \alpha &> 3f^+ + 2 \frac{f^{+2}}{\lambda} \end{aligned} \tag{18}$$

Then the observer scheme is stable, and the states of the observer in Eq. (15) (\hat{x}_1, \hat{x}_2) converges to the true state (x_1, x_2) in Eq. (6) in finite time.

Proof: by similar way with the Ref. [13], we can verify that the observer states converge to the true states in finite time.

5 Fault Tolerant Control Scheme

5.1 Fault Reconstruction

Consider the estimation error as expressed in Eq. (16) converges to zero. Eq. (16) can be written as

$$\begin{aligned}\dot{\tilde{x}}_1 &\equiv 0 \\ \dot{\tilde{x}}_2 &= \varepsilon + \phi(x_1, x_2, u) - \alpha \text{sign}(\tilde{x}_1) \equiv 0\end{aligned}\quad (19)$$

Notice that $d(x_1, x_2, \hat{x}_2, u) = f(x_1, x_2, u) - f(x_1, \hat{x}_2, u) = 0$ due to the estimation states converge to the true states ($\hat{x}_2 = x_2$). Then the equivalent output injection (EOI) is defined as

$$z_{eq} = \alpha \text{sign}(\tilde{x}_1) = \varepsilon + \phi(x_1, x_2, u) \quad (20)$$

Theoretically, the equivalent output injection is the result of an infinite switching frequency of the discontinuous terms $\alpha \text{sign}(\tilde{x}_1)$ that is so-called chattering. To eliminate this high frequency chattering, we use a low pass filter has the form

$$a \dot{\bar{z}}_{eq}(t) + \bar{z}_{eq}(t) = z_{eq}(t) \quad (21)$$

where a is the filter time constant.

After the filtration, we have

$$z_{eq} = \bar{z}_{eq} + \gamma \quad (22)$$

where every elements of \bar{z}_{eq} is the filtered version of z_{eq} and γ is the difference caused by filtration process.

Nevertheless, as it is shown in [13]:

$$\lim_{a \rightarrow 0} \bar{z}_{eq}(t) = z_{eq}(t), \quad a \in \mathfrak{R}$$

Now, once the information of the equivalent output injection is available, it is possible to reconstruct the fault function by mean of the following expression

$$\bar{z}_{eq} = \hat{\phi}(x_1, x_2, u) + \hat{\varepsilon} \quad (23)$$

where $\hat{\phi}(x_1, x_2, u)$, $\hat{\varepsilon}$ are the estimation of $\phi(q, \dot{q}, \tau)$ and ε , respectively. Because $\phi(x_1, x_2, u) \gg \varepsilon$, hence the fault estimation can be approximated by $\bar{z}_{eq} \approx \hat{\phi}(x_1, x_2, u)$.

5.2 Fault Tolerant Control Scheme

The key problem of the FTC scheme is online reconfiguration of the controller based on the obtained fault estimation. The FTC architecture is illustrated in Fig. 1 that consists of three parts: 1) the CTC where the real joint velocity measurement is replaced by the estimated velocity; 2) the estimated uncertainties compensating; 3) the estimated fault compensating.

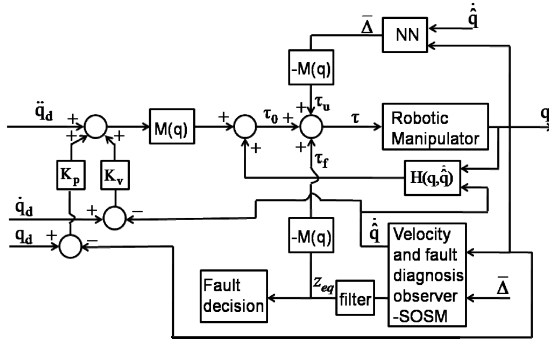


Fig. 1. Structure of fault tolerant control scheme

The fault tolerant control scheme is design as

$$\tau = \tau_0 + \tau_u + \tau_f \tag{24}$$

where $\tau_0 = M(x_1)(\ddot{q}_d + K_V(\dot{q}_d - \hat{x}_2) + K_P(q_d - x_1)) + H(x_1, \hat{x}_2)$, τ_u and τ_f are used to compensate for the modeling uncertainties and faults, respectively. They are defined as follows:

$$\tau_u = -M(x_1)W_0\sigma(V_0\bar{x}(t)) \tag{25}$$

and

$$\tau_f = -M(x_1)\bar{z}_{eq} \tag{26}$$

6 Simulation Results

In order to verify the effectiveness of the proposed algorithm, its overall procedure is simulated for a PUMA560 robot where the first three joints are used. Its explicit dynamic model and its parameter are given in Ref. [14]. The uncertainties that include friction, small joint backlash are given by

$$F_u = \begin{bmatrix} \dot{q}_1 + \sin(3q_1) \\ 1.1\dot{q}_2 + 0.8\sin(2q_2) \\ 0.8\dot{q}_3 + 1.1\sin(q_3) \end{bmatrix} \tag{27}$$

In this simulation, the sliding gains are chosen $\alpha = 12$, $\beta = 8$. To estimate the uncertainties, a three layer neural network has 20 neurons in hidden layer, the tuning law was given in eqs. (10) and (11) with $\rho_1 = \rho_3 = 5$ and $\rho_2 = \rho_4 = 0.03$. First in normal operation, Fig. 2 shows the uncertainty estimation performance of the neural network observer. We see that the neural network observer obtain good uncertainty approximation.

To verify the performance of the FTC, we assumed a fault $\phi = [0, 6, 4 \sin(t)]^T$ occurs in the second and third joints at $t = 10 \text{ s}$. The time histories of joint angles are given in Fig. 3 without fault estimation compensating ($\tau = \tau_0 + \tau_u$). The fault signal is reconstructed from the EOI is given by $z_{eq} = \alpha_i \text{sign}(\tilde{x}_{li})$ using a low pass filter with $a = 0.002$, the result is shown in Fig. 4. From this figure, we see that the fault are correctly detected, isolated and identified. The mismatch between the EOI of sliding mode and the fault function inevitably come from the filtration and model uncertainty estimation error. Fig. 5 shows the performance of the fault tolerant control scheme under the presence of the fault ϕ . Comparison of Fig. 5 with Fig. 3 shows that the tracking performance is improved due to its model self-correction capability.

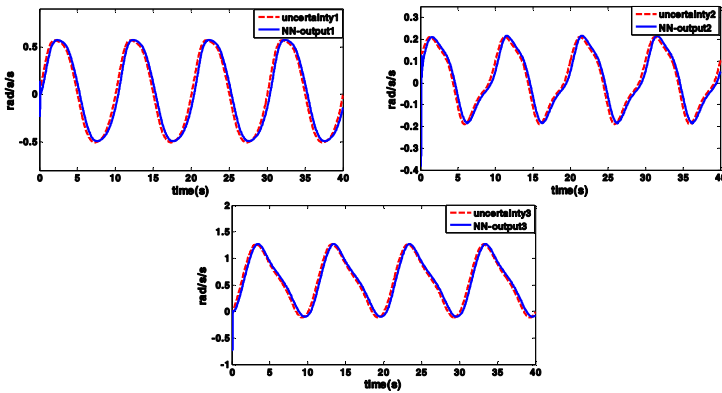


Fig. 2. Uncertainties and their estimation using neural network

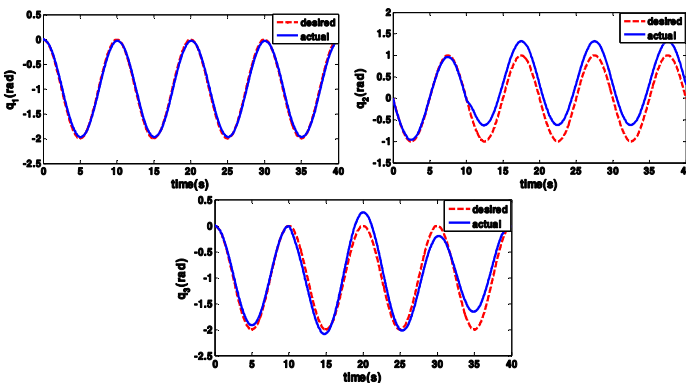


Fig. 3. Joint angles of the robot manipulator when fault ϕ occurs without fault estimation compensating

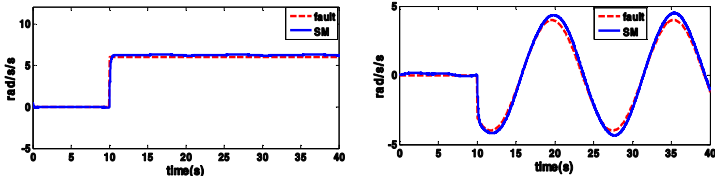


Fig. 4. Fault identification on the second and third joints

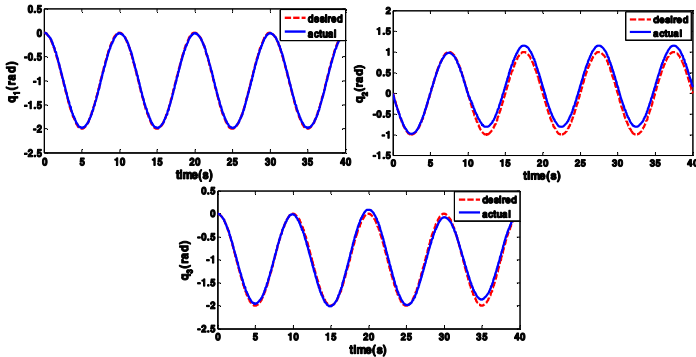


Fig. 5. Joint angles of the robot manipulator when fault ϕ occurs with fault estimation compensating

7 Conclusions

An active fault tolerant control scheme for uncertainty robot manipulator using only position measurement has been presented in this paper. Here, the uncertainty observer for estimation of modeling uncertainties and the second-order sliding mode observer for the fault diagnosis are investigated to show their effectiveness. In addition, a fault tolerant control based on a modification of computed torque control is suggested for better tracking performance in the system. The results of computer simulations for a 3-DOF PUMA560 robot verify the effectiveness of the proposed strategy.

Acknowledgements. The authors would like to express financial supports from Korean Ministry of Knowledge Economy both under Human Resources Development Program for Convergence Robot Specialists and under Robot Industry Core Technology Project.

References

1. Gertler, J.J.: Survey of Model-Based Failure Detection And Isolation in Complex Plants. IEEE Control Systems Magazine 8, 3–11 (1988)
2. Vemuri, A.T.: Neural-Network-Based Robust Fault Diagnosis In Robotic Systems. IEEE Transaction on Neural Networks, 1410–1420 (1997)

3. Song, Q., Yin, L.: Robust Adaptive Fault Accommodation for A Robot System Using A Radial Basic Function Neural Network. *International Journal of System Science* 32, 195–204 (2001)
4. Christopher, E., Spurgeon, S.K., Patton, R.J.: Sliding Mode Observers for Fault Detection And Isolation. *Automatica* 36, 541–553 (2000)
5. Brambilla, D., Capisani, L.M., Ferrara, A., Pisu, P.: Fault Detection for Robot Manipulators Via Second-Order Sliding Modes. *IEEE Transactions on Industrial Electronics* 55, 3954–3963 (2008)
6. Davila, J., Fridman, L., Levant, A.: Second-Order Sliding-Mode Observer for Mechanical Systems. *IEEE Transactions on Automatic Control* 50, 1785–1789 (2005)
7. Jiang, J., Yu, X.: Fault Tolerant Control Systems: A Comparative Study Between Active And Passive Approaches. *Annual Reviews in Control* 36, 60–72 (2012)
8. Tao, G., Chen, S., Joshi, S.M.: An Adaptive Actuator Failure Compensation Controller Using Output Feedback. *IEEE Transactions on Automatic Control* 47, 506–511 (2002)
9. Huang, S.N., Tan, K.K.: Fault Detection, Isolation, And Accommodation Control in Robotic Systems. *IEEE Transaction on Automation Science and Engineering* 5, 480–489 (2008)
10. Van, M., Kang, H.-J., Suh, Y.-S., Shin, K.-S.: A Robust Fault Diagnosis And Accommodation Scheme for Robot Manipulators. *International Journal of Control, Automation, and Systems* 11, 377–388 (2013)
11. Kim, Y.H., Lewis, F.L.: Neural Network Output Feedback Control of Robot Manipulators. *IEEE Transactions on Robotics and Automation* 15, 301–309 (1999)
12. Abdollahi, F., Talebi, H.A., Patel, R.V.: A Stable Neural Network-Based Observer with Application To Flexible-Joint Manipulators. *IEEE Transactions on Neural Networks* 17, 118–129 (2006)
13. Moreno, J.A., Osorio, M.: A Lyapunov Approach to Second-Order Sliding Mode Controllers and Observers. In: *47th IEEE Conference on Decision and Control*, pp. 2856–2861 (2008)
14. Armstrong, B., Khatib, O., Burdick, J.: The Explicit Dynamic Model And Inertial Parameters of The PUMA 560 Arm. Stanford University, Artificial Intelligence Laboratory, IEEE (1986)

Fusion of Vision and Inertial Sensors for Position-Based Visual Servoing of a Robot Manipulator

Tran Minh Duc¹ and Hee-Jun Kang^{2,*}

¹ Graduate School of Electrical Engineering, University of Ulsan,
680-749, Ulsan, South Korea
ductm.ctme@gmail.com

² School of Electrical Engineering, University of Ulsan,
680-749, Ulsan, South Korea
hjkang@ulsan.ac.kr

Abstract. This paper presents a new technique to combine measurements from both a vision system and an inertial sensor mounted on a robot tip when the robot moves towards an object. The proposed scheme is designed based on a combination of the relative pose (position and orientation) of vision system with information of the inertial sensor (acceleration and angular velocity) to get final relative pose. In addition, a Kalman filter is used to handle asynchronous information from those sensors. The proposed scheme not only can increase the robustness of the pose estimation but also can smooth the velocity control of robot. A computer simulation results are shown to verify the effectiveness of the proposed algorithm.

Keywords: Visual Servoing, Visual-Inertial Servoing, Multi-rate System, Sensor Fusion, Relative Pose Estimation.

1 Introduction

Visual and inertial perception is an important combination in nature. It is believed that humans easily manipulate in any new environment by using visual and inertial information from eyes and the inner ear [1]. Recently, many researchers have investigated to combine the inertial sensor (IMU) with the vision together in the robot system [1, 10-11, 13-15]. The combined vision and IMU system determines an orientation and navigation solution, ego motion and structure sensing. The approaches mentioned above make the robot more intelligent and flexible.

In visual servoing, many approaches have been proposed to estimate the 3D relative position and orientation by using a single image view [2-5]. Normally, the approaches are designed based on two methods. In the first method, people use specific image preprocessing tasks to extract the image feature location measurements. They are then combined with a known object CAD descriptions to estimate the pose of the object with respect to camera frame. Then, the estimated pose

* Corresponding author.

can be composed with the object of desired pose transformation to find out the relationship between the current and the desired pose for effective dynamic tracking control. In the second method, the Kalman filter uses these measurements for the implicit solution of the photogrammetric equation relating to the image plane feature point coordinates to give optimal pose estimation for real-time tracking control [3-4]. However, these methods still have some limitations such as the effect of measurement noise, the camera intrinsic parameter calibration errors, the target model errors, the stability of the closed-loop system, the steady-state error [5]. All of them make the pose estimation very sensitive. It is known that the convergence of position based visual servoing may be prevented by the loss of stability of pose estimation. In order to overcome this problem, several approaches have been proposed [6-8]. Gans et al. [8] proposed a switching between position and image based visual servoing. Deng [6] has proposed the use of the steady state error as a measure of sensitivity of visual servoing. However, this algorithm needs a long trajectory to obtain a good result. Another approach considers the outliers of the image data to increase the robustness of the control law [7].

The purpose of this study is to prevent the sensitivity of the pose estimation by using fusion sensors. The combination of vision and inertial measurement offers several advantages. Firstly, visual sensing is very accurate at low velocities while inertial sensors can track fast motions at high frequency. The combination enables the construction of a sensor system which is able to estimate the pose during slow motions and fast motions. Secondly, in mathematical physics, vision sensor provides only 2D projection of the 3D object when it is projected onto the real world image. Therefore, it is just able to get actual position with derivative order 0. In another case, the inertial sensor can provide gyroscopes and accelerometers with order 1 and order 2 [1]. From analysis mentioned above, the resulting of combination of vision and inertial sensors is much better compared to that of only using vision or inertial sensors.

The most common researches of fusion sensors focused on motion tracking and registration for augmented reality applications. Chai [13] presented an architecture to combine vision and inertial sensor in which the multiple – model extended kalman filters estimate head estimation. Armesto [14] and H. Rebbinder [15] have improved the ego motion estimation and compared the performance of EKF and UKF. However, their jobs applied for only mobile robots (augmented reality). Recently, Nguyen et al. [10-11] has presented a new method to combine visual and inertial for visual servoing. In his study, an inertial Jacobian is newly defined that is combined with the conventional Jacobian for the proposed visual servoing algorithm. However, the limitation of this method is that it has only been developed in image based visual servoing. In this paper, a new approach to combine the inertial sensor with vision for PBVS is presented. The system set-up is an end-effector mounted camera together with an inertial sensor and a Kalman filter is used to handle asynchronous inputs from visual and inertial sensor. By combining the vision sensor with inertial sensor, the resulting scheme preserves the advantages of both techniques, such as the accurate estimation of the pose during slow motion of the vision sensor and during fast motions with the inertial sensor. The pose estimation increases the robustness and stability of the control law.

2 Fusion of Vision and Inertial Sensor

The goal of the fusion filtering is to estimate the camera pose parameters from the measurement of the vision and inertial sensor.

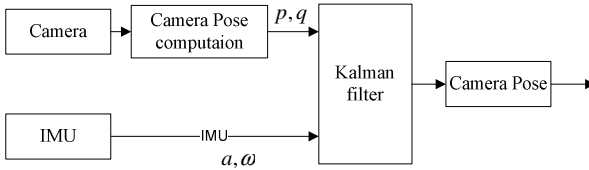


Fig. 1. Fusion of vision and inertial sensor

In this scheme, the inertial sensors provide the acceleration information in three coordinate axes and the three angular velocities of these directions. The vision information is taken from a camera. Then, the image feature location measurements are extracted by image preprocessing. The results are then combined with a known object CAD descriptions to estimate the pose of the object with respect to camera frame. Finally, a Kalman filter is used to handle asynchronous inputs from visual and inertial sensor which has a different measurement equation and dynamic equation. Whenever one of the sensors sends new data, the corresponding measurement equation is used. The output of this filter is the pose estimation of the camera with respect to an initial coordinate system.

2.1 Filter Equation

The state estimation is designed based on a discrete-time non-linear state-space description of the head pose.

$$x(k+1) = f[x(k), w(k)] \tag{1}$$

$$y(k) = h[x(k)] + v(k) \tag{2}$$

where f is the dynamic description of the system, w is process noise, h is the mapping of the state to the output equation and v is the measurement noise.

The state vector $x(k)$, the system noise $w(k)$, the measurement $y(k)$, and the measurement noise $v(k)$ are defined as follows.

$$x(k) = [p^T, v^T, a^T, b_a^T, q^T, \omega^T]_k^T$$

$$y(k) = [a_m^T, w_m^T, p_m^T, q_m^T]_k^T$$

$$w(k) = [j^T, \alpha^T, b^T]_k^T$$

$$v(k) = [v_{am}^T, v_{\omega m}^T, v_{pm}^T, v_{qm}^T]^T_k$$

The state consists of position and orientation. Position is represented by Cartesian position $p(k)$, velocities $v(k)$ and accelerations $a(k)$. The orientation is represented by quaternion $q(k)$ and angular velocities $w(k)$. Biases of the acceleration measurements b_a^T . The measurement vector is formed with accelerations a_m^T and angular velocities w_m^T measurements from the inertial sensor and the Cartesian positions p_m^T and quaternion q_m^T measurements from the vision sensor. The system noises are considered by Jerks $j(k)$, angular accelerations $\alpha(k)$ and velocity biases $b'(k)$.

The dynamic equation can be separated into two parts:

$$x_t(k+1) = f_t[x(k), w(k)] \tag{3}$$

$$x_r(k+1) = f_r[x(k), w(k)] \tag{4}$$

where $x_t(k)$ describes the translational part and $x_r(k)$ describes the rotation part.

$$x_t(k) = [p^T, v^T, a^T, b_a^T]^T_k$$

$$x_r(k) = [q^T, \omega^T]^T_k$$

The translational dynamic equations have the form

$$p(k+1) = p(k) + \Delta T.v(k) + \frac{\Delta T^2}{2} a(k) + \frac{\Delta T^3}{6} j(k) \tag{5}$$

$$v(k+1) = v(k) + \Delta T.a(k) + \frac{\Delta T^2}{2} j(k) \tag{6}$$

$$\begin{aligned} a(k+1) &= a(k) + \Delta T.a'(k) \\ &= a(k) + \Delta T(j(k) + \alpha(k) \times v(k) + \omega(k) \times v(k)) \end{aligned} \tag{7}$$

$$b_a(k+1) = b_a(k) + \Delta T b'(k) \tag{8}$$

where ΔT is the sampling period.

The translational part of the state vector $x_t(k)$ is composed of equations (5)-(8) and can be computed as.

$$\begin{aligned}
 x_t(k+1) = & \begin{bmatrix} I_{3 \times 3} & \Delta T I_{3 \times 3} & \frac{\Delta T^2}{2} & 0_{3 \times 3} \\ 0_{3 \times 3} & I_{3 \times 3} & \Delta T I_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} \end{bmatrix} \begin{bmatrix} p \\ v \\ a \\ b_a \end{bmatrix}_k + \begin{bmatrix} 0_{3 \times 3} \\ 0_{3 \times 3} \\ \Delta T(\omega \times a) \\ 0_{3 \times 3} \end{bmatrix}_k \\
 & + \begin{bmatrix} 0_{3 \times 3} \\ 0_{3 \times 3} \\ \Delta T(a \times v) \\ 0_{3 \times 3} \end{bmatrix}_k + \begin{bmatrix} \frac{\Delta T^3}{6} I_{3 \times 3} \\ \frac{\Delta T^2}{2} I_{3 \times 3} \\ \Delta T I_{3 \times 3} \\ 0_{3 \times 3} \end{bmatrix} j(k) + \begin{bmatrix} 0_{3 \times 3} \\ 0_{3 \times 3} \\ 0_{3 \times 3} \\ \Delta T I_{3 \times 3} \end{bmatrix}_k b'(k)
 \end{aligned} \tag{9}$$

The rotational dynamic equations have the form

$$\omega(k+1) = \omega(k)I_{3 \times 3} + \alpha(k).\Delta T.I_{3 \times 3} \tag{10}$$

$$\dot{q} = \frac{1}{2} q \times \Omega \tag{11}$$

The quaternion integration equation is approximated by a discrete equation, see Suh et al. [9].

$$\begin{aligned}
 q(k+1) = & (I + \frac{3}{4}\Omega(k)\Delta T - \frac{1}{4}\Omega(k-1).\Delta T - \frac{1}{6}[\omega(k)]_2^2 \Delta T^2 \\
 & - \frac{1}{24}\Omega(k)\Omega(k-1)\Delta T^2 - \frac{1}{48}[\omega(k)]_2^2 \Omega(k)\Delta T^3)q(k)
 \end{aligned} \tag{12}$$

Where

$$\Omega = \begin{bmatrix} 0 & -\omega_x & -\omega_y & -\omega_z \\ \omega_x & 0 & \omega_z & -\omega_y \\ \omega_y & -\omega_z & 0 & \omega_x \\ \omega_z & \omega_y & -\omega_x & 0 \end{bmatrix} \tag{13}$$

The state presentation in (1) and (2) is always the same. The output equations are linear with the state defined and can be written as

$$y(k) = h[x(k)] + v(k) = Hx(k) + v(k) \tag{14}$$

where

$$H = \begin{bmatrix} 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 3} & I_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 4} \\ 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & I_{3 \times 4} \\ I_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 4} \\ 0_{4 \times 3} & 0_{4 \times 3} & 0_{4 \times 3} & 0_{4 \times 3} & I_{4 \times 3} & 0_{4 \times 4} \end{bmatrix} \tag{15}$$

2.2 Multi-rate Filters

The goal of the multi-rate filtering is to estimate the camera pose from the measurements of the vision and inertial (IMU) sensors. Because the vision and inertial sensors have different sample rates, the Kalman filters are designed to handle asynchronous input from two sensors as shown in Figure below.

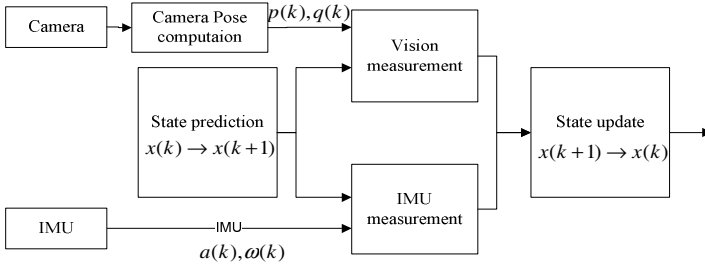


Fig. 2. Fusion filter framework

The idea of the multi-rate filter is to build the output vector containing only those measurements. There are two independent input channels in prediction module (vision and inertial sensors). At each instant frame time, we first predict the filter state based on the measurement sample, if it is available. Second, the multi-rate sample is implemented so that it generates a size varying output y , and then corrects the prediction using the observed vision and inertial measurements, given by equation (4) and (5).

The recursive EKF algorithm consists of two major parts of prediction and estimation as follows.

State prediction:

$$\hat{x}_k^- = A\hat{x}_{k-1} \tag{16}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{17}$$

Kalman gain update:

$$K_k = P_k^- H^T (R + HP_k^- H^T)^{-1} \tag{18}$$

Estimation update:

$$\hat{x}_k = \hat{x}_k^- + K_k (z_k - H\hat{x}_k^-) \tag{19}$$

$$P_k = P_k^- - K_k H_k P_k^- \tag{20}$$

3 Position Based Visual Servoing

Fig. 3 describes the operation of the proposed algorithm for position based visual servoing. The scheme is designed based on a combination of information of vision system with information of inertial sensor through a kalman filter to get final relative pose s . Here, Inertial sensor is used to get accelerometer and gyros information, and vision sensor is used to get the relative pose. Then, the Kalman filter is used to handle asynchronous information of the inputs. The final relative pose will be used to design the control law, as the following.

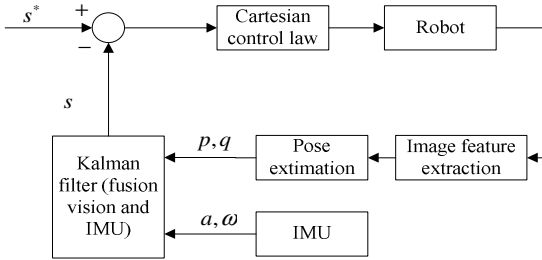


Fig. 3. Position-based visual servoing structure

It is convenient to consider three coordinate frames: the current camera frames \mathfrak{S}_c , the desired camera frame \mathfrak{S}_{c^*} , and a reference frame \mathfrak{S}_o attached to the object.

${}^c_o t$: The coordinate vector of (\mathfrak{S}_o) origin with respect to current camera frame.

${}^{c^*}_o t$: The coordinate vector of (\mathfrak{S}_o) origin with respect to desired camera frame.

${}^c_b R$: Rotation matrix from frame \mathfrak{S}_a origin to frame \mathfrak{S}_b .

${}^c v \equiv [v_c, \omega_c]$: Cartesian velocity considered in camera coordinate frame \mathfrak{S}_c .

The PBVS (position based visual servoing) scheme can be designed by using $s = ({}^c_o t, {}^c_o \Theta)$, $s^* = 0$, $e = s$. The conventional controller is satisfied, $\dot{e} = \lambda e$. In this case, the interaction matrix related to e is given by

$$\dot{e} = \dot{s} = \begin{bmatrix} {}^c_o \dot{t} \\ {}^c_o \dot{\Theta} \end{bmatrix} = \begin{bmatrix} {}^c_o R & 0 \\ 0 & T^{-1} {}^c_o R \end{bmatrix} \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} \tag{21}$$

Where the matrix T is a relationship between angle set velocity and the angular velocity of the camera. It is derived from Euler representation [12] such as

$$T(\Theta) = \begin{bmatrix} 0 & -\sin \varphi & \cos \varphi \cos \theta \\ 0 & \cos \varphi & \sin \varphi \cos \theta \\ 1 & 0 & -\sin \theta \end{bmatrix} \tag{22}$$

Follow exponential control method $\dot{e} = \lambda e$ chaumette[3], the required control command is expressed as

$${}^c V_c = \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} = -\lambda \begin{bmatrix} ({}^c R)^T \cdot {}^c t & o \\ o & ({}^c R)^T T ({}^c \Theta) \cdot {}^c \Theta \end{bmatrix} \quad (23)$$

4 Simulation Results

To verify the effectiveness of the proposed algorithm, a PBVS simulator is used and then it is compared with conventional method. All the simulations are performed in Matlab/Simulink. In this simulation, the object is fixed and four landmarks are set up in the same plane. The initial Euler angles are selected as $[0; \frac{\pi}{4}; \frac{\pi}{6}]$, while destination angles are chosen $[0;0;0]$. The initial position is set at $[0;0;3]$, while destination position is set at $[0;0;1.5]$. The random noise in image measurement is also added in this simulation.

$${}^c P = ([0;0;3], [0; \frac{\pi}{4}; \frac{\pi}{6}]); {}^o P = ([0;0;1.5], [0;0;0]); \quad (24)$$

The Fig. 4a shows the resultant trajectories of 4 landmarks in the image plane for visual-inertial PBVS. The red ones are start points and the green ones are destination points. The Fig. 4b shows a good convergence of feature parameters error. It helps to prove the control quality of improved PBVS algorithm.

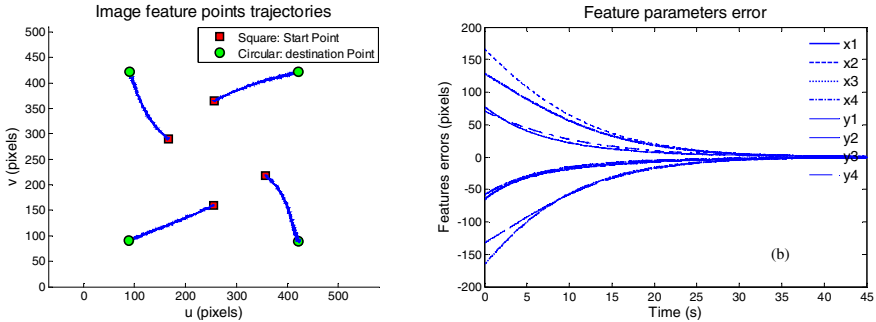


Fig. 4. a) Trajectories of landmarks in Position based Visual Servoing, b) Convergence of visual feature error in PBVS

The Fig. 5a shows the control velocity in case of fusion sensors. Fig. 5b shows the control velocity for conventional PBVS. From these figures, it is clearly shown that the velocity of the conventional method is not smooth for systems with random noise in measurement image. By employing the proposed algorithm, the velocity control is greatly smooth. This indicates that the quality of the proposed fusion sensors is higher than with that of conventional method for PBVS.

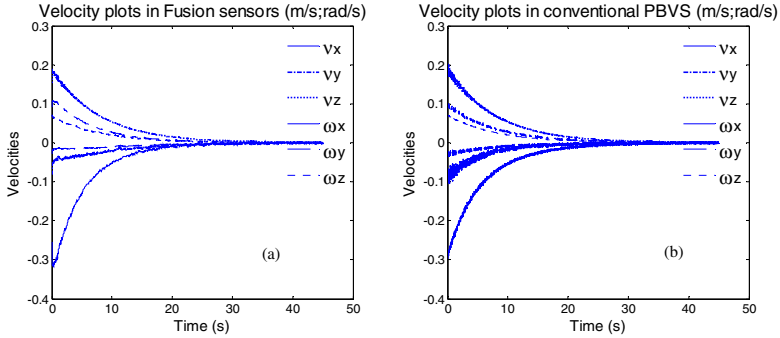


Fig. 5. a) The evolution of the six components of the camera velocity in Fusion sensors for PBVS. b) The evolution of the six components of the camera velocity in conventional PBVS.

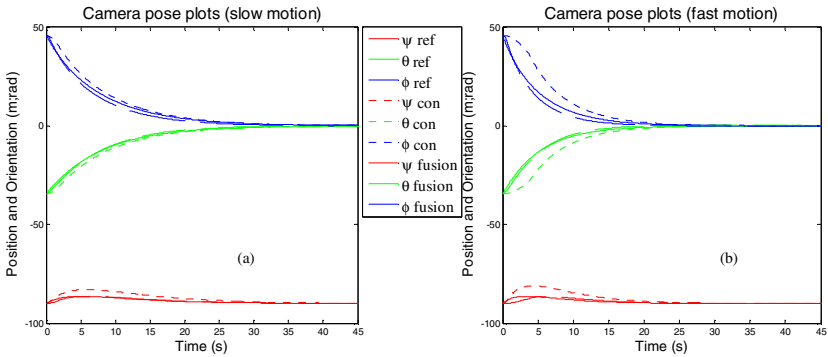


Fig. 6. a) Camera orientation at slow motion, b) Camera orientation at fast motion

The Fig. 6a and 6b show the comparison between the conventional method and propose method in terms of camera orientation estimation. From these figures, we see that the camera pose estimation of the proposed method is more accurate than the conventional method in both slow motion (a) and fast motion (b).

5 Conclusion

This paper presented the fusion of vision and inertial sensors in position-based visual servoing. The most important advantage of the system is that it increases the robustness and stability of robot control. This combination makes the robot more intelligent and flexible. The fusion sensors compensate the weakness of each other to increase the reliability of the sensors system. The proposed scheme not only can increase the robustness of the pose estimation but also smoother the velocity control of robot. The simulation results are shown to verify the effectiveness of the proposed algorithm.

Acknowledgements. This research was supported by the 2011 Research Fund of University of Ulsan.

References

1. Corke, P., Lobo, J., Dias, J.: An Introduction to Inertial and Visual Sensing. *The International Journal of Robotics Research* 26, 519 (2007)
2. Chaumette, F., Hutchinson, S.: Visual Servo Control. Part I: Basic Approaches. *IEEE Robotics & Automation Magazine* (December 2006)
3. Wilson, W.J., Hulls, C.W., Bell, G.S.: Relative End-Effector Control Using Cartesian Position Based Visual Servoing. *IEEE Transactions on Robotics and Automation* 12(5) (October 1996)
4. Janabi-Sharifi, F.V., Marey, M.: A Kalman-Filter-Based Method for Pose Estimation in Visual Servoing. *IEEE Transactions on Robotics* 26(5) (October 2010)
5. Kragic, D., Christensen, H.I.: Measurement Errors in Visual Servoing. In: *International Conference on Robotics & Automation* (April 2004)
6. Deng, L., Wilson, W.J., Janabi-Sharifi, F.: Characteristics of Robot Visual Servoing Methods And Target Model Estimation. In: *Proceedings of the 2002 IEEE International Symposium on Intelligent Control, Vancouver, Canada* (2002)
7. Comport, A., Pressigout, M., Marchand, E., Chaumette, F.: A visual servoing control law that is robust to image outliers. In: *Proceedings of the 2003 IEEE/RSJ Conference on Intelligent Robots and Systems, Las Vegas* (2003)
8. Gans, N.R., Hutchinson, S.A.: An Asymptotically Stable Switched System Visual Controller for Eye in Hand Robots. In: *Proceedings of the 2003 IEEE/RSJ Conference on Intelligent Robots and Systems, Las Vegas* (2003)
9. Suh, Y.S.: Orientation Estimation Using a Quaternion-Based Indirect Kalman Filter with Adaptive Estimation of External Acceleration. *IEEE Transactions on Instrumentation and Measurement* 59(12) (2010)
10. Nguyen, H.Q.P., Kang, H.J., Suh, Y.S., Ro, Y.S.: A Visual-Inertial Servoing Method for Tracking Object with Two Landmarks and an Inertial Measurement Unit. *International Journal of Control, Automation and System* 19(2), 317–327 (2011)
11. Nguyen, H.Q.P., Kang, H.J., Suh, Y.S., Elle, O.J.: A Platform Stabilization Algorithm Based on Feed-forward Visual-Inertial Servoing. *International Journal of Precision Engineering and Manufacturing* 13(4), 517–526 (2012)
12. Craig, J.J.: *Introduction to Robotics Mechanics and Control*, 3rd edn.
13. Chai, L., Hoff, W.A., Vincent, T.: 3-D Motion and Structure Estimation Using Inertial Sensors and Computer Vision for Augmented Reality
14. Armesto, L., Chroust, S., Vincze, M., Tornero, J.: Multi-rate Fusion with Vision And Inertial Sensors. In: *International Conference on Robotics & Automation* (2004)
15. Rebbinder, H., Ghosh, B.: Multi-rate Fusion with Vision And Inertial Data. In: *International Conference on Multi-Sensor Fusion and Integration for Intelligent Systems* (2001)

Efficient 3D Reconstruction for Urban Scenes

Weichao Fu, Lin Zhang*, Hongyu Li, Xinfeng Zhang, and Di Wu

School of Software Engineering, Tongji University, Shanghai, China
cslinzhang@tongji.edu.cn

Abstract. Recently, researchers working in the fields of computer graphics and computer vision have shown tremendous interests in reconstructing urban scenes. For this task, the acquisition of the 3D point clouds is the first step, for which scans are usually widely utilized. Nevertheless, on seeing the potential drawbacks of scans, in this paper, we propose a novel urban scene reconstruction system based on the Multi-View Stereo (MVS). Given a set of calibrated photographs, we first generate point clouds using an existing MVS algorithm, and then reconstruct the sub-structures that often regularly repeat in urban buildings. Finally, we recover the entire architectural models through an automatic growing algorithm of the sub-structures in dominant directions. Experimental results on regular urban buildings show the practicality and high efficiency of the proposed reconstruction method.

Keywords: Urban Scene, Multi-View Stereo, sub-structures.

1 Introduction

Reconstruction of urban scenes is attracting increasing attention these days, motivated by ambitious applications that aim to build digital copies of real cities (e.g., Microsoft Virtual Earth 3D and Google Earth 3D) [1]. It is a very complex problem with a lot of research history, open problems, and possible solutions.

Usually, the first step for an urban scene reconstruction system is the acquisition of 3D point clouds and the state-of-the-art methods, such as [2], utilized laser scans for this step. However, as an approach for 3D point clouds acquisition from urban scenes, laser scans has some inherent drawbacks with respect to the cost and the easiness for distribution. In addition, the main difficulty of laser scans of large-scale urban environments is the data quality. Large distances between the scanner and the scanned objects reduce the precision and imply higher level of noise. The obtained point clouds usually exhibit significant missing data. As a result, in this paper, we propose a new urban scene reconstruction scheme based on another 3D point clouds acquisition approach, multi-view stereo (MVS). Compared with laser scans, MVS has the advantage of low cost to get data sources (e.g., Internet photo sharing sites like Flickr and Google Images). And for the urban scene, the quality of point clouds obtained by MVS is usually satisfactory. According to a quantitative evaluation on the Middlebury benchmark, a patch-based

* Corresponding author.

MVS method [3] outperforms all others for most of the datasets. Hence, we make use of this PMVS algorithm to recover point clouds.

Urban landscapes exhibit a high degree of self-similarity and redundancy [1]. This characteristic of urban scenes is not a chance occurrence, but is demonstrated universally across countries and cultures. Such large scale repetitions arise from aesthetics, build-ability, manufacturing ease, etc. Also urban buildings are mostly comprised of flat or near-planar faces because of functional requirements and constraints. Consequently, the presence of such characteristics suggests opportunities for simplifying the reconstruction task.



Fig. 1. Images of urban scenes with texture-poor but highly structured surfaces are increasingly ubiquitous on the Internet

In this paper, we propose a 3D urban scene reconstruction method based on exploration of properties of architectural scenes. Our method is inspired by recent works of [2], which assemble an architectural model over a 3D point cloud by interactive operations. The key idea is to replace the smoothness prior used in traditional methods with priors that are more appropriate for urban buildings. Our approach operates as follows. Given a set of calibrated photographs, we first generate point clouds using an existing MVS algorithm, and then reconstruct the sub-structures that often regularly repeat in urban buildings. At this step, regularity and self-symmetry properties of urban buildings [1] are explored. Finally, we recover the entire architectural models through an automatic growing algorithm of the sub-structures in dominant directions. Our method, exploits properties of urban scenes, offers the following advantages. 1) It is remarkably robust to lack of texture, and able to model flat painted walls shown in Fig. 1(b). 2) It can work with poor quality point clouds generated from photographs due to occlusion, as shown in Fig. 1(c). 3) It produces remarkably clean and simple models as outputs.

This paper is arranged as follows. In section 2 we review the PMVS algorithm and preprocess the point clouds generated by PMVS. In section 3 we reconstruct the sub-structures that often regularly repeat in urban buildings and in section 4 we recover the entire architectural models through an automatic growing algorithm of the sub-structures. Experimental results are presented in section 5, followed by conclusions in section 6.

2 MVS Preprocessing

PMVS algorithm presented in [3] is implemented as a match, expand, and filter procedure, starting from a sparse set of matched keypoints, and repeatedly expanding these before using visibility constraints to filter away false matches. Given a set of calibrated photographs, the first step of our method is to use the PMVS algorithm [3] to generate a set of oriented 3D points (positions and normal). We retain only high-confidence points in textured areas. Associated with each point P_i , several attributes can be extracted, including the 3D location, a surface normal N_i , a set of visible images V_i , and a photometric consistency score (normalized cross correlation) $C(P_i)$.

2.1 Outliers Removing

Although the PMVS algorithm proposed in [3] enforces local photometric consistency and global visibility constraints to remove outliers, the point clouds generated by PMVS still have sparse outliers which will corrupt the estimation of local point cloud attributes such as surface normal or curvature changes. We solve these irregularities by performing a statistical analysis on each point's neighborhood, and trimming those which do not meet a certain criteria. Our sparse outlier removal scheme is based on the distribution of distances from a point to its neighbors in the PMVS point clouds. For each point, we compute the mean distance from it to all its neighbors. By assuming that the resulted distribution is Gaussian with a mean and a standard deviation, all points whose mean distances are outside an interval defined by the global distances mean and standard deviation can be considered as outliers and trimmed from the point clouds.

An example of the outliers-removal is shown in Fig. 2. Fig. 2(a) shows the original point cloud obtained by PMVS, while Fig. 2(b) shows the result of outliers-removal. We set the number of nearest neighbors as 50 for mean distance estimation.

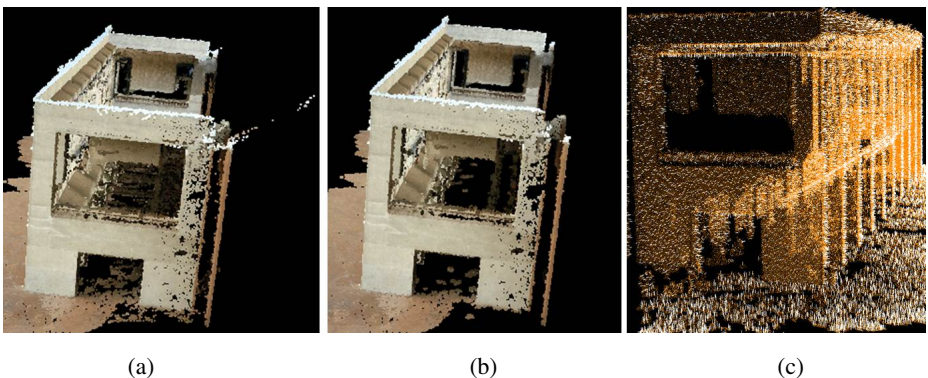


Fig. 2. (a) The original point cloud obtained by PMVS. (b) The resultant one after outliers removing. (c) The filtered point cloud with normals.

2.2 Estimating Normals and Curvature

The PMVS algorithm in [3] optimizes the normal of each patch by simply minimizing the photometric discrepancy score, and constraining the number of degrees of freedom to two (yaw and pitch). However, the normal optimized in PMVS is not good enough for urban scene, because of the large distance between building surface and the optical center of the camera, and the initialization of PMVS algorithm works poorly for the large scale urban buildings.

Though many different normal estimation methods exist, the one we used in this paper is one of the simplest, and is formulated as follows. The problem of determining the normal to a point on the surface is approximated by the problem of estimating the normal of a plane tangent to the surface, which in turn reduces to an analysis of the eigenvectors and eigenvalues of a covariance matrix created from the nearest neighbors of the query point. More specifically, for each point P_i , we assemble the covariance matrix C as follows:

$$C = \frac{1}{k} \sum_{i=1}^k (P_i - \bar{P})(P_i - \bar{P})^T, C \cdot \vec{V}_j = \lambda_j \cdot \vec{V}_j, j \in \{0, 1, 2\} \quad (1)$$

where k is the number of point neighbors considered in the neighborhood of P_i , \bar{P} represents the 3D centroid of the nearest neighbors, λ_j is the j -th eigenvalue of the covariance matrix, and \vec{V}_j is the j -th eigenvector. The point cloud with normals is shown in Fig. 2(c). The surface curvature change is estimated from the eigenvalues as:

$$\frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2}, \text{ where } \lambda_0 < \lambda_1 < \lambda_2 \quad (2)$$

2.3 Extracting Dominant Directions

We could require the dominant axes of buildings in an automatic growing algorithm of the sub-structures. Hence, we employ a simple greedy algorithm using the normal estimates N_i recovered above (See [7, 8, 9] for similar ideas). We compute a histogram of normal directions over a unit sphere, subdivided into 1000 bins. We then set the first dominant axis \vec{d}_1 to the average of the normals within the largest bin. Next, we find the largest bin which is in the range 80 to 100 degrees away from \vec{d}_1 and set the second dominant axis \vec{d}_2 to the average normal within that bin. Finally, the third dominant axis \vec{d}_3 is computed in the same way. We allow for some deviation from orthogonality to compensate for possible errors in normal estimation and to handle urban building that itself is not composed of exactly orthogonal planes.

3 Sub-structure Reconstruction

While in recent years many techniques have been developed to detect repeated parts in models [5, 11, 12, 13] and regularity directly on 3D geometry [6], in this paper, we attempt to learn the repetitions pattern directly from high level user guidance for the existence of occlusion or missing data. Due to the generally poor quality of generated data, we argue that some level of user interaction is required to understand the semantics of the building model. Hence, at the beginning of this section, we select the points that imply the sub-structures by our interactive tool. Then, a region growing algorithm is utilized for simplifying the reconstruction of the sub-structures.

3.1 Region Growing Algorithm

The purpose of the region growing algorithm is to merge the points that are close enough in terms of the smoothness constraint. Thereby, the output of this algorithm is a set of clusters, and each cluster has a set of points that are considered to be a part of the same smooth surface.

We first sort the points by their curvature value. It is necessary because the region begins its growth from the point that has the minimum curvature value. The reason for this is that the point with the minimum curvature is located in the flattest area. Then, as long as there are unlabeled points in the cloud, we pick up the point with minimum curvature value and start the growth of the region. First, the picked point is added to a set called seeds. Then, for every seed point we find its neighboring points, and test the angle between the normal of the current seed point and its neighbors' normals. If the angle is less than a threshold θ , the current point is added to the current region. After that every neighbor is tested for the curvature value. If the curvature is less than a threshold δ , then this point is added to the seeds. Finally, current seed is removed from the seeds, and if the seed set becomes empty which means that a certain region has grown, we repeat the process from the beginning. The overall pseudo code for this growing algorithm is given in Table 1.

3.2 Reconstruction

After the region growing algorithm, we employ an efficient RANSAC method for point-cloud shape detection in each region of the sub-structures. The RANSAC algorithm we used is presented in [9]. It decomposed the point cloud into a concise, hybrid structure of inherent shapes and a set of remaining points. The method can detect planes, spheres, cylinders, cones and tori. However, because of functional requirements and constraints, urban buildings are mostly comprised of flat or near-planar faces. Consequently, the RANSAC method with plane model detection is usually sufficient. Also, the region with too little points can be rejected or determined by high level user guidance which RANSAC model can be adopted. Eventually, the sub-structure is generated by enforcing alignment among the RANSAC models. This step will resolve the inconsistencies we wish to repair, e.g., model intersections, small gaps, and other forms of misalignments.

Table 1. The pseudo code for the region growing algorithm

<p><i>Input:</i> Point cloud = $\{P\}$, Point normals = $\{N\}$, Point curvatures = $\{C\}$, Neighbor finding function $\Omega(\cdot)$, Curvature threshold δ, Angle threshold θ.</p> <p><i>Output:</i> Region list $\{R\}$.</p> <hr/> <p>$R \leftarrow \phi$, Available points list $\{A\} \leftarrow \{P_1, \dots, P_n\}$</p> <p>While $\{A\}$ is not empty do</p> <p style="padding-left: 2em;">Current region $\{R_c\} \leftarrow \phi$, Current seeds $\{S_c\} \leftarrow \phi$</p> <p style="padding-left: 2em;">Point with minimum curvature in $\{A\} \rightarrow P_{\min}$</p> <p style="padding-left: 2em;">$\{S_c\} \leftarrow \{S_c\} \cup P_{\min}$, $\{R_c\} \leftarrow \{R_c\} \cup P_{\min}$, $\{A\} \leftarrow \{A\} \setminus P_{\min}$</p> <p style="padding-left: 2em;">for $i = 0$ to size $(\{S_c\})$ do</p> <p style="padding-left: 4em;">Find nearest neighbors of current seed point $\{B_c\} \leftarrow \Omega(S_c\{i\})$</p> <p style="padding-left: 4em;">for $j = 0$ to size $(\{B_c\})$ do</p> <p style="padding-left: 6em;">Current neighbor point $P_j \leftarrow B_c\{j\}$</p> <p style="padding-left: 6em;">If $\{A\}$ contains P_j and $\cos^{-1}(N\{S_c\{i\}\}, N\{B_c\{j\}\}) < \theta$ then</p> <p style="padding-left: 8em;">$\{R_c\} \leftarrow \{R_c\} \cup P_j$</p> <p style="padding-left: 8em;">$\{A\} \leftarrow \{A\} \setminus P_j$</p> <p style="padding-left: 6em;">If $C\{P_j\} < \delta$ then</p> <p style="padding-left: 8em;">$\{S_c\} \leftarrow \{S_c\} \cup P_j$</p> <p style="padding-left: 6em;">end if</p> <p style="padding-left: 4em;">end if</p> <p style="padding-left: 2em;">end for</p> <p style="padding-left: 2em;">Add current region to global region list $\{R\} \leftarrow \{R\} \cup \{R_c\}$</p> <p>end while</p>
--

4 Automatic Continuation

Our sub-structure automatic continuation algorithm is inspired by recent works in [2]. However, in [2], some essential interactive operations are utilized for creating complex structures through the propagation of a repeated structure. In our case, the automatic continuation of sub-structure is constrained along the dominant directions computed in Section 2, which can minimize the required user interaction.

Also, in [2], an optimization which balances between the data-fitting and contextual forces is adopted. We continue this trend and present our data-fitting term and contextual term below. Given a sub-structure S and the point cloud, the data-fitting term, also known as the fitting error is computed by measuring the one-sided Euclidean distance from the sub-structure to points [14]. The contextual sub-structure \bar{S} required to define the contextual force is a previously positioned sub-structure in a reconstruction sequence. The contextual term is defined as the sum of two terms: the interval term and the alignment term. The interval term $I(S, \bar{S})$ measures how well

the interval length between S and \bar{S} agrees with the expected interval length due to regularity constraints. The alignment term $A(S, \bar{S})$ measures how well corresponding edges of S and \bar{S} align. These two terms are normalized independently and weighted equally to form the contextual term.

The automatic growing algorithm is achieved by finding an optimal linear transformation T^* , consisting of translation and scaling of the sub-structure S , to minimize a weighted sum of data-fitting and contextual terms,

$$T^* = \arg \min_{\langle T \rangle} \left\{ \omega \cdot D(T(S), P) + (1 - \omega) \cdot C(T(S), \bar{S}) \right\} \quad (3)$$

where ω is a weight which balances between the data-fitting and contextual forces. The choice of the weight ω should correlate with an assessment of data quality and contextual regularity in the urban building model to be reconstructed. In our method, the default setting of ω is 0.5. Nevertheless, the user can manually adjust it according to data quality and contextual regularity perceived by them.

5 Experimental Results

The proposed reconstruction method was tested on both a publicly available data set: Hall [4] and our own images of Building F in Tongji University (taken by cell phone

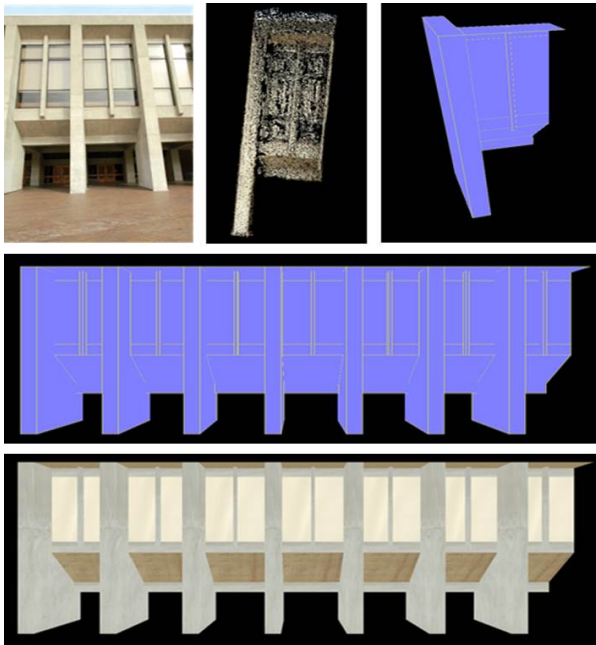


Fig. 3. Top left: a target image. Top middle: the point cloud of the sub-structure. Top right: sub-structure reconstruction. Middle: sub-structure growing among the dominant direction. Bottom: the reconstructed model with texture mapping.

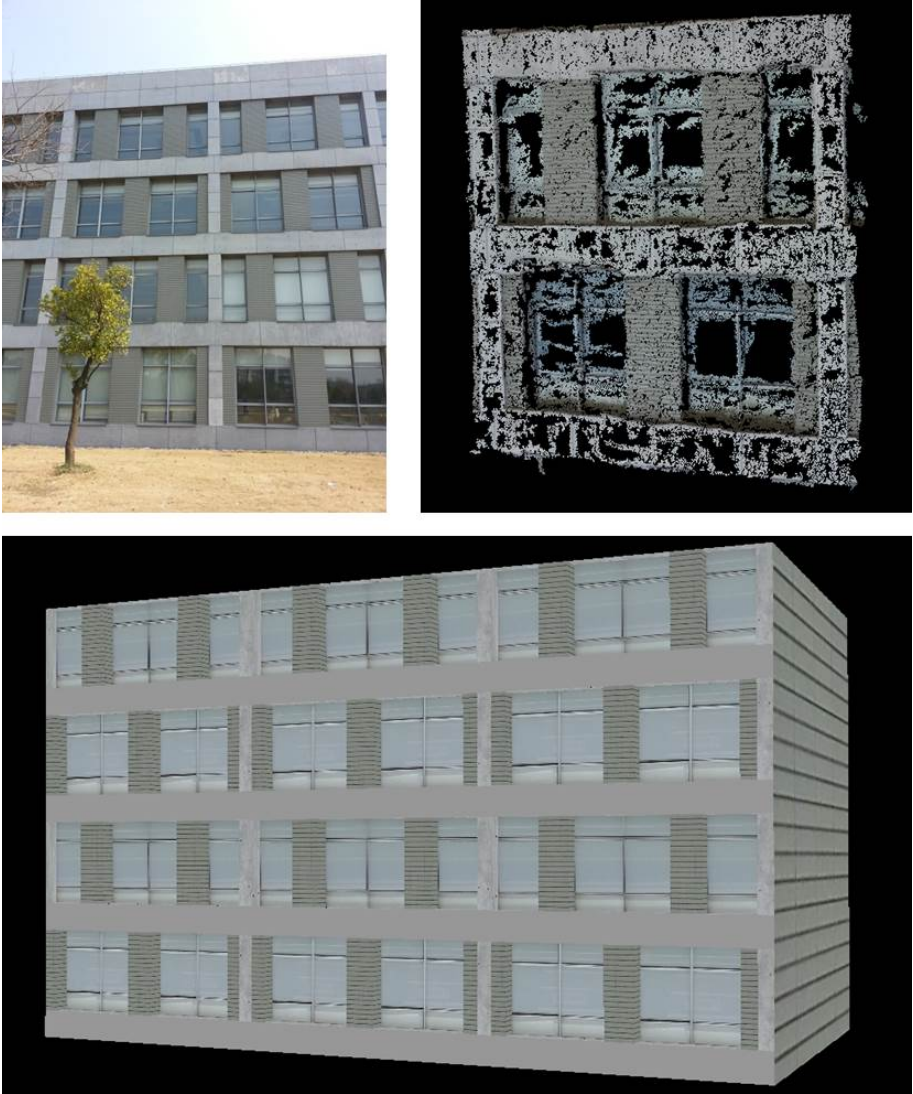


Fig. 4. Top left: a target image. Top right: the point cloud of the sub-structure. Bottom: the reconstructed model with texture mapping.

camera). The Hall dataset consists of 61 images which are used for PMVS algorithm to generate point cloud. This results in a total of 452393 points. In our case we use only the high-confidence points, thus, we have 435935 points after the MVS preprocessing step. We set $\theta = 0.122$ and $\delta = 1.0$ for the thresholds used in region growing algorithm. The interval term used in automatic continuation is initialized as the length of the sub-structure's projection among the dominant direction. Our own

Building F dataset consists of 15 images taken by cell phone camera, and some of the images have occlusion as shown in top left of Fig. 4. The reconstruction results are shown in Fig. 3 and Fig. 4, which illustrate the convenience and robustness of our proposed method.

6 Conclusions and the Future Work

In this paper, we have demonstrated a Multi-View stereopsis based 3D urban scene reconstruction method that utilizes the large scale repetitions and self-similarities of urban buildings. Our method is remarkably robust to buildings lack of texture, and also able to work with poor quality point clouds by exploiting properties of urban scenes. In future work, we plan to incorporate any parametric primitive that can be detected using standard RANSAC in our sub-structure reconstruction step. This may lead to a better model and exhibit the diversity of modern urban scenes.

Acknowledgement. This work is supported by the Fundamental Research Funds for the Central Universities under grant no. 2100219033, the Natural Science Foundation of China under grant no. 61201394, and the Innovation Program of Shanghai Municipal Education Commission under grant no. 12ZZ029.

References

1. Zheng, Q., Sharf, A., Wan, G., Li, Y., Mitra, N.J., Cohen-Or, D., Chen, B.: Non-local Scan Consolidation for 3D Urban Scenes. *ACM Transactions on Graphics* 29, 94 (2010)
2. Nan, L., Sharf, A., Zhang, H., Cohen-Or, D., Chen, B.: SmartBoxes for Interactive Urban Reconstruction. *ACM Transactions on Graphics* 29, 93 (2010)
3. Furukawa, Y., Ponce, J.: Accurate, dense, and robust Multi-View stereopsis. In: *CVPR*, pp. 1–8 (2007)
4. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Manhattan-world Stereo. In: *CVPR*, pp. 1422–1429 (2009)
5. Liu, J., Musialski, P., Wonka, P., Ye, J.: Tensor Completion for Estimating Missing Values in Visual Data. In: *ICCV*, pp. 2114–2121 (2009)
6. Pauly, M., Mitra, N.J., Wallner, J., Pottmann, H., Guibas, L.J.: Discovering Structural Regularity in 3D Geometry. *ACM Transactions on Graphics* 27, 43 (2008)
7. Coorg, S., Teller, S.: Extracting Textured Vertical Facades from Controlled Close-Range Imagery. In: *CVPR*, pp. 625–632 (1999)
8. Pollefeys, M., Nister, D., Frahm, J.M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.J., Merrell, P.: Detailed Real-Time Urban 3D Reconstruction from Video. In: *IJCV* 2008, pp. 143–167 (2008)
9. Werner, T., Zisserman, A.: New Techniques for Automated Architectural Reconstruction from Photographs. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002, Part II. LNCS*, vol. 2351, pp. 541–555. Springer, Heidelberg (2002)
10. Schnabel, R., Wahl, R., Klein, R.: Efficient RANSAC for Point-Cloud Shape Detection. *Computer Graphics Forum* 26, 214–226 (2007)

11. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial And Approximate Symmetry Detection for 3D Geometry. *ACM Transactions on Graphics* 25, 560–568 (2006)
12. Korah, T., Rasmussen, C.: Spatiotemporal Inpainting for Recovering Texture Maps of Occluded Building Facades. *IEEE Transactions on Image Processing* 16, 2262–2271 (2007)
13. Pauly, M., Mitra, N.J., Wallner, J., Pottmann, H., Guibas, L.J.: Discovering Structural Regularity in 3D Geometry. *ACM Transactions on Graphics* 27, 43 (2008)
14. Nan, L., Xie, K., Sharf, A.: A Search Classify Approach for Cluttered Indoor Scene Understanding. *ACM Transactions on Graphics* 31, 137 (2012)

Combining Edge and One-Point RANSAC Algorithm to Estimate Visual Odometry

Van-Dung Hoang, Danilo Cáceres Hernández, and Kang-Hyun Jo

Graduated School of Electrical Engineering, University of Ulsan, Korea
{dungvanhoang,danilo}@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. In recent years, classical structure from motion based SLAM has achieved significant results. Omnidirectional camera-based motion estimation has become interested researchers due to the larger field of view. This paper proposes a method to estimate the 2D motion of a vehicle and mapping by using EKF based on edge matching and one point RANSAC. Edge matching based azimuth rotation estimation is used as pseudo prior information for EKF predicting state vector. In order to reduce requirement parameters for motion estimation and reconstruction, the vehicle moves under nonholonomic constraints car-like structured motion model assumption. The experiments were carried out using an electric vehicle with an omnidirectional camera mounted on the roof. In order to evaluate the motion estimation, the vehicle positions were compared with GPS information and superimposed onto aerial images collected by Google map API. The experimental results showed that the method based on EKF without using prior rotation information given error is about 1.9 times larger than our proposed method.

Keywords: Omnidirectional camera, edge feature matching, one-point RANSAC, motion and mapping.

1 Introduction

Autonomous vehicle and mapping are an important research in various applications, e.g. path planning and mapping, intelligent transport, and surveillance systems. Some progresses have been made in this field during the last few years. In recent years, many methods have been developed for navigation, visual odometry[1],[2]. Some researchers used vision only, e.g. monocular camera, stereo camera, and catadioptric camera for visual odometry or used only electro-magnetic devices, e.g., Inertial Measurement Unit (IMU), wheel odometer, and laser rangefinder sensors. Other researches combined both electro-magnetic devices and vision systems for improving accuracy and absolute transformation estimation.

The related works are separated into several categories. In the first approach group, the early researches proposed vision odometry method using a single perspective camera [3, 4]. Other authors proposed methods using the binocular camera [5, 6]. Because of angle of view limitation, some authors used the omnidirectional camera. Typically, the omnidirectional vision based odometry systems were presented in [7-10]. The basic

principle of these approaches is feature point correspondence and epipolar geometry constraint. The main obstacle of this approach is accumulative error over time. The global trajectory is acceptable if vehicle moves in indoor environments, or in special outdoor environments. The trajectory will be diverged compared with the ground truth when vehicle moves on long distance without any prior information. This is also a challenge for the incremental methods of visual odometry or Simultaneous localization and mapping (SLAM). Moreover, the scale of trajectory process is ambiguous if a monocular vision is used. In the second approach group integrate multi-magnetic sensors in the system[11],[12]. Normally, GPS receiver is used for global position, IMU or wheel odometers are used for the local position estimation. These methods often yield the correct results on the large-scale scene. However, GPS is not always correct under environment occurring high obstacle-high building in urban. This is also a challenge for localization, navigation without vision systems. In recent years, the combination of the two methods mentioned above is considered as the solution to overcome these disadvantages. Some authors, typically in [13],[14],[15] proposed methods based on the fusion of the vision system and the electro-magnetic devices. The results were significantly improved. Due to limited of precision of sensors and additional ambient noise, there is location and mapping estimation errors. Over time, the errors will also be accumulated when the vehicles move in the large scale scene. Therefore the final global trajectory and map will be diverged and distorted. The system can yield the accurate results in short distance of movements or integrating with information from Global Positioning System (GPS) receiver. In recent years, several filter technique were applied to deal with this problem, e.g. (Extended) Kalman filter, particle filters.

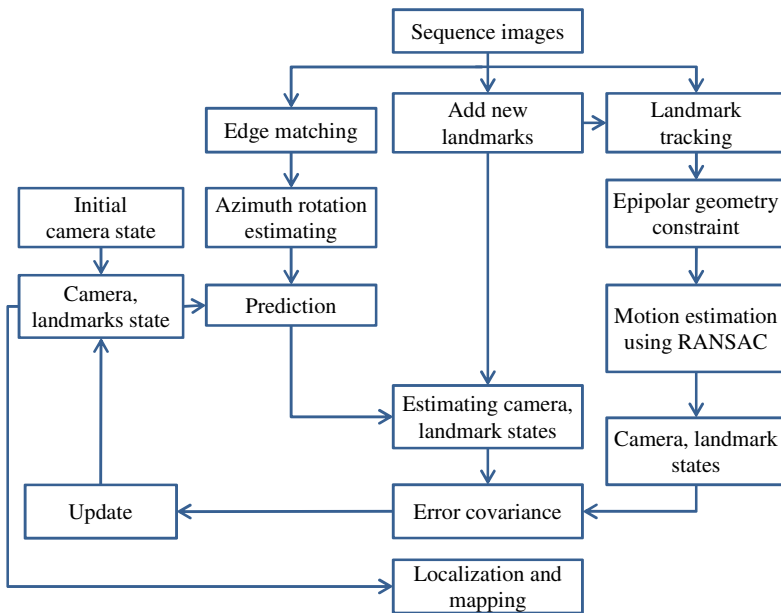


Fig. 1. The method proposed for vehicle motion estimation and mapping

Nowadays, the omnidirectional vision has become interesting in application for mobile robot/vehicle and intelligent transport system [1, 7-10]. The main advantage of omnidirectional image is the direction of objects directly computed and tracking in wide scene. This is the reason omnidirectional are often used in mobile robotics. However, one drawback of omnidirectional image that the surrounding scene is transformed to a ring shaped, low resolution and distortion for reconstruction.

In this paper, a planar motion estimation method is proposed using monocular omnidirectional camera based on Extended Kalman Filter (EKF). The vehicle motion parameters consist of rotation and translation factors. Vision information is used for vehicle rotation estimation with assumption constant velocity for translation approximation. Edge matching based azimuth rotation estimation is used as pseudo prior information for EKF prediction vehicle motion and mapping. In EKF observation step, one point RANSAC is used to estimating mapped landmarks. The overview of the proposed method is represented in Fig. 1.

2 Motion Estimation

2.1 Geometry Constraint

The visual odometry system is composed of consecutive image pair constraints. Those constraints are analyzed directly from the epipolar constraint using the essential matrix. Fig. 3(a) shows image point correspondences (m, m') project on mirror respective two projection rays r, r' from the focus points of the hyperboloid mirror to P . The relative constraint can be described as follows:

$$r'^T E r = 0 \tag{1}$$

where the essential matrix E is defined as $E=[T]_x R$. The matrix $[T]_x$ is a skewed symmetric matrix of translation, and R is rotation matrix.

$$R = R_z R_y R_x = \begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\varphi - \sin\alpha\cos\varphi & \sin\alpha\sin\varphi + \cos\alpha\sin\beta\cos\varphi \\ \sin\alpha\cos\beta & \cos\alpha\cos\varphi + \sin\alpha\sin\beta\sin\varphi & \sin\alpha\sin\beta\cos\varphi - \cos\alpha\sin\varphi \\ -\sin\beta & \cos\beta\sin\varphi & \cos\beta\cos\varphi \end{bmatrix} \tag{2}$$

$$[T]_x = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \tag{3}$$

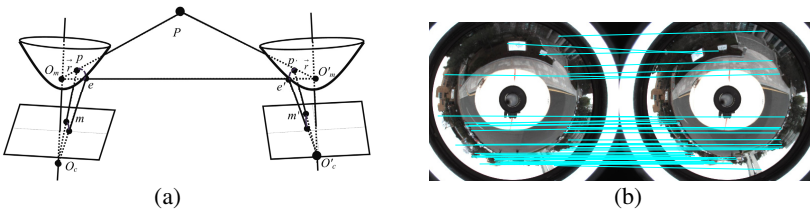


Fig. 2. (a) Epipolar geometry and (b) corresponding points

The corresponding r and r' vector of the two image sequence are computed based on calibration omnidirectional camera. Corresponding points between consecutive omnidirectional images are extracted by the SIFT method [16]. Then the method in [17] is used to construct the reprojection ray r and r' from the center of the mirror to a point in 3D space. The point P is the intersection of r and r' . In particle however, due to limited precision of processing r and r' not intersect, P can be estimated at the point which nearest both of lines of two ray r and r' . The problem solution is solving (1) to compute translation and rotation.

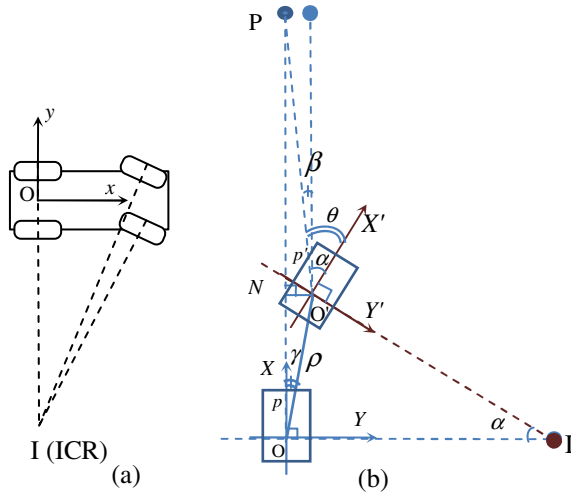


Fig. 3. Related camera poses under car-like structure constrain

2.2 One-Point for Motion Estimation

In order to minimize the number of correspondence point requirement, the vehicle moves under assumption planar ground and following car-like structure model [8, 18]. Therefore, the relation of successive poses can be represented by an rotation α and a translation $T=[t_x, t_y, 0]^T$. The car-like structure is described by Ackermann steering geometry in Fig. 3(a). The circular center is known as the instantaneous centers of rotation (ICR). The number of discrete rotations around different center of rotations composes the final motion. The camera was mounted on the roof at rear-wheel axis of vehicle. The distance from camera position to the rear wheel axis is considered to solve using the method in [8]. While the vehicle moves along an ICR in model, O and O' lie on the same the circular center I on the occasion that OX and $O'X'$ are tangents of the circle. Therefore, two parts of the translation are with the relative angle $\gamma=\alpha/2$. The relation of the two camera poses is represented by rotation and translation matrix equations:

$$R = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad T = \rho \begin{bmatrix} \cos(\alpha/2) \\ \sin(\alpha/2) \\ 0 \end{bmatrix} \quad (4)$$

Therefore, the essential matrix is

$$E = \begin{bmatrix} 0 & 0 & \sin(\alpha/2) \\ 0 & 0 & -\cos(\alpha/2) \\ \sin(\alpha/2) & \cos(\alpha/2) & 0 \end{bmatrix} \quad (5)$$

The solution is solved by one point correspondence for constructing corresponding ray $r = [x_r, y_r, z_r]^T$ and $r' = [x_{r'}, y_{r'}, z_{r'}]^T$, and then substitution (5) into (1) for solution:

$$\alpha = 2 \tan^{-1} \frac{z_r y_{r'} - y_r z_{r'}}{x_r z_{r'} + z_r x_{r'}} \quad (6)$$

2.3 Rotation Estimation Based on Edge Matching

Relation between vehicle rotation and appearance scenes at the front of vehicle is used to estimation vehicle rotation as Fig. 3(b). Two OX and $O'X'$ axes are coincided with head direction of vehicle at two consecutive positions. The landmark P appears in the images of current and next camera poses respect to p and p' . Assume that the landmark P was projected into the current image on OX - head direction. The angle α is computed based on a disparity between p' and p . The angle θ is rotation angle between two camera poses. A residual angle β is a disparity between α and θ angle, $\alpha = \theta + \beta$. The angle β is rate of NO' and NP , $\beta = \tan^{-1}(NO'/NP)$. The translation interval ρ is computed by approximation velocity of previous step. The landmark in front of vehicle is large enough for vehicle motion. Therefore, NO' is negligible against NP , β angle is extremely small, α can be approximated to θ . This property is used for estimating vehicle rotation as description in next section.

In omnidirectional image, scenes on two sides of camera mirror change in opposite directions when the camera moves straight. The landmarks in front and rear of vehicle mirror change quite slowly and separately onto two mirror's sides. Landmarks are uniform changed under camera rotation. Similarly, in [9, 19], omnidirectional camera is used as a visual compass for rotation estimation. The corresponding edges of landmarks in front and rear scenes of the vehicle are matched by using the Chamfer matching method [20]. The Chamfer edge matching is implemented on panoramic image, which is unwrapped from omnidirectional image by using the method in [21].

Firstly, the templates region in the front scenes of the vehicle of the current frame and the searching region in the next frame are extracted respective Fig. 4(a), and Fig. 4(b). To reduce time consuming and occurred obstacles, the searching regions are restricted in field of view of about 30° around the front scene of vehicle, which is shown in the red color window in Fig. 4(b). The edges of both templates and searching regions are extracted by Canny edge detection method.

Secondly, the distance transformations (DT) of both templates and searching regions are computed. Initially, all edge pixels are set to zero value and the remains non-edge pixels are set to infinity. The integer approximation of Euclidian distance was used to compute the DT by followed expression:

$$D_{i,j}^k = \min(D_{i-1,j}^{k-1} + 2, D_{i-1,j-1}^{k-1} + 3, D_{i,j-1}^{k-1} + 2, D_{i+1,j-1}^{k-1} + 3, D_{i+1,j}^{k-1} + 2, D_{i+1,j+1}^{k-1} + 3, D_{i,j+1}^{k-1} + 2, D_{i-1,j+1}^{k-1} + 3, D_{i,j}^{k-1}) \quad (7)$$

where $D_{i,j}^k$ is the DT value of position (i, j) from nearest edge, at k^{th} iteration. This process iterates until the value of each pixel is not change.

Thirdly, the template and the searching region are matched by superimpose and sliding the template on the searching region. The mean square of subtraction pixel values between DT of template and DT of searching region is computed. It is called the edge distance. This value is zero in the perfected matching case. Finally, the rotation angle between successive camera poses is computed as the equation below:

$$\alpha = \frac{180}{\pi R} (x_{i+1} - x_i) \quad (8)$$

where R is the radius of the out bounding circle respect to width of panoramic image. x_i and x_{i+1} are the center matching locations on horizontal axis in the current image and the next image, respectively.

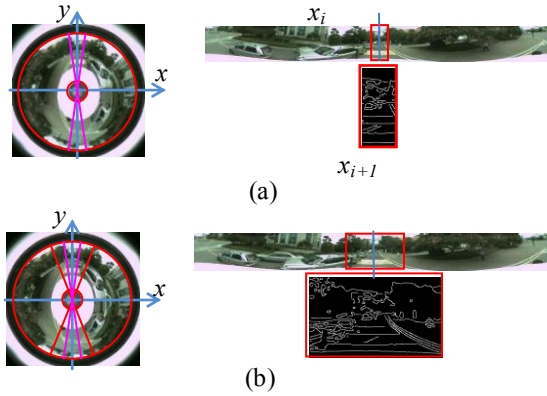


Fig. 4. Candidate regions for matching, (a) the template window extract from current frame, (b) the searching region extracts from the next frame

3 Combining Edge and One-Point for Visual Odometry

Nowadays, EKF has become a standard technique of nonlinear dynamic system estimation. This section will not be reviewed here, for details of EKF in SLAM see also [22][23][24]. This paper focuses in the case of combination pseudo prior information rotation based on edge matching for predicting vehicle motion and one-point RANSAC measurement for updating localization and mapping.

The map is a state vector stacking vehicle and environment landmarks $X=[C,L]^T$, where C is the vehicle state and L set of n landmarks (L_1, \dots, L_n) states of the current

map. The number of landmark n is increase over time. The map is represented in Cartesian coordinate by a Gaussian stochastic under zero-mean white noise assumption, estimated mean \widehat{X} and error covariance P as $X \sim N(\widehat{X}, P)$. The vehicle state consists of position, motion directional and linear velocity $C=[x_C, y_C, \alpha_C, v_C]^T$. The landmark is Euclidean point representing by location in 3D $L_i=[x_{L_i}, y_{L_i}, z_{L_i}]^T$. The detail of algorithm for estimating motion and mapping is described in Fig.5.

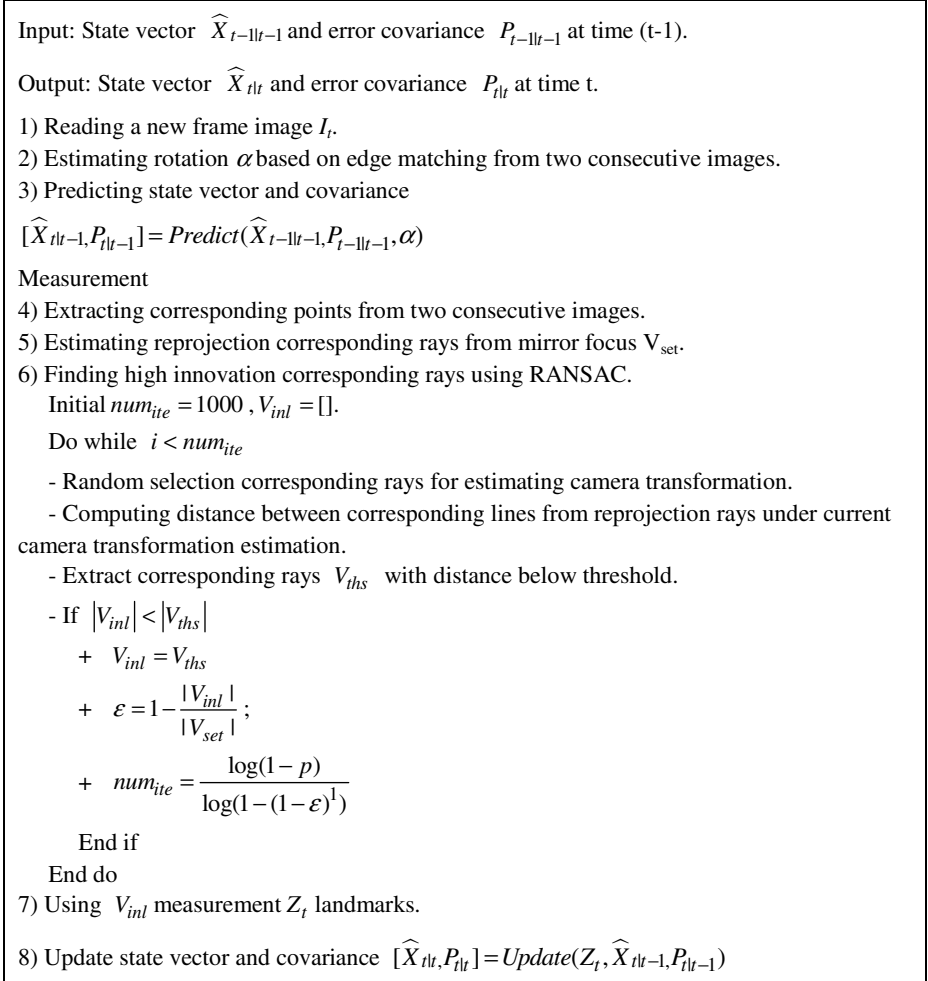


Fig. 5. The algorithm for motion estimation based on EKF using edge and one-point RANSAC

4 Experiments

The experiments were performed in loop closure travel. The consecutive images collected by an omnidirectional camera system constituted classical perspective camera



Fig. 6. Vehicle equipped with sensors: Omnidirectional camera and GPS receiver

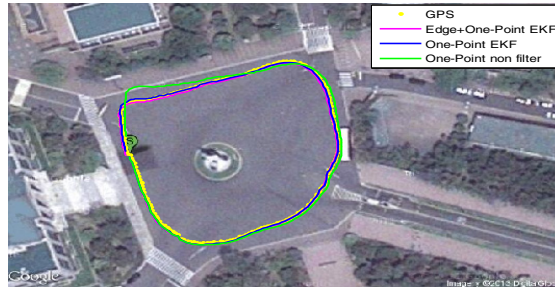


Fig. 7. Plot final trajectory of vehicle on aerial Google image

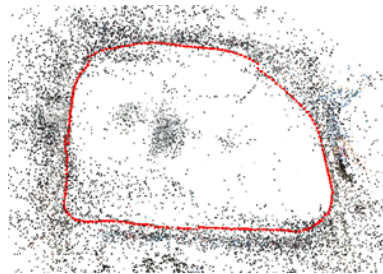


Fig. 8. Plotting proposed method based trajectory and 3D point clouds

Table 1. Comparison error at starting and ending of closer loop travel

<i>Method</i>	<i>Meter</i>	<i>%</i>
Edge+ one-point EKF	1.4	0.67
One-point EKF	2.6	1.24
One-point non filter	9.07	4.34

and hyperboloid mirror. It was mounted on an electric vehicle as shown in Fig. 5. For edge matching, front and back regions of omnidirectional image were unwrapped into panoramic image. The omnidirectional image direction was defined at the first frame, collinear with the vehicle head direction. Parameters of omnidirectional image are (960×1280) pixel resolution, (460,646) pixel center point, 150 pixel inner circular radius, 475pixel outer circular radius. The camera system was calibrated by toolbox [17].

The proposed method using edge matching based rotation and one-point RANSAC EKF was compared with one-Point RANSAC EKF, and one-point RANSAC without filter (Table 1). The SIFT technique [16] was used for feature point detection and

matching. The SIFT features are stable for matching, but expensive for consumption cost. For initial velocity, GPS information was used to compute absolute translation. In order to evaluate, GPS information was used as ground truth. Fig. 6 shows trajectory result of 209m travelled around rotary with 206 frames. Table 1 shows the comparison error at starting and ending positions of closer loop travel. In order to compare with ground truth and plot on Google map image, the absolute translation of between sequence frame is extracted from GPS are used for all methods. The EKF based localization using prior rotation information for predicting state of vehicle given higher precision than direct predicting vehicle motion by approximating previous step. The method only using one-point RANSAC without filter gave highest divergent trajectory due to miss estimation at any frame step is not rectified which used for next step.

5 Conclusions

This paper presented a method to estimate planar motion of vehicle based on EKF using monocular omnidirectional camera. From omnidirectional image, the prior rotation information is extracted by edge matching, which used for EKF predicting state vector of vehicle. Some advantages were pointed out through the arguments and the experiments. The edge matching based visual compass works accurately in complicated scenes and difference lighting conditions of outdoor environments. The robustness and rapidity of this approach ensure be applicable in real time systems. Under car-like structure motion model, one-point RANSAC is used for estimation vehicle motion for EKF observation. This system also compensates for one of them miss matching. Our future works focus on improving the method for real time, and high accuracy based on sensors fusion systems for odometry and SLAM under 3D motion.

Acknowledgements. This work was supported by The Ministry of Knowledge Economy, Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H1502-12-1002).

References

1. Fraundorfer, F., Scaramuzza, D.: Visual Odometry: Part II: Matching, Robustness, Optimization, Applications. *IEEE Robotics & Automation Magazine* 19, 78–90 (2012)
2. Grasa, O.G., Civera, J., Montiel, J.M.M.: EKF Monocular SLAM with Relocalization for Laparoscopic Sequences. In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 4816–4821 (2011)
3. Nistér, D., Naroditsky, O., Bergen, J.: Visual Odometry for Ground Vehicle Applications. *Journal of Field Robotics* 23, 3–20 (2006)
4. Royer, E., Lhuillier, M., Dhome, M., Lavest, J.-M.: Monocular Vision for Mobile Robot Localization and Autonomous Navigation. *Int. J. Comput. Vis.* 74, 237–260 (2007)
5. García, D.V., Rojo, L.F., Aparicio, A.G., Castelló, L.P., García, O.R.: Visual Odometry through Appearance- and Feature-Based Method with Omnidirectional Images. *Journal of Robotics* 13 (2012)

6. Konolige, K., Agrawal, M., Solà, J.: Large-Scale Visual Odometry for Rough Terrain. In: Kaneko, M., Nakamura, Y. (eds.) *Robotics Research*, vol. 66, pp. 201–212 (2011)
7. Gandhi, T., Trivedi, M.: Parametric Ego-Motion Estimation for Vehicle Surround Analysis Using an Omnidirectional Camera. *Machine Vision and Applications* 16, 85–95 (2005)
8. Scaramuzza, D.: 1-Point-RANSAC Structure from Motion for Vehicle-Mounted Cameras by Exploiting Non-holonomic Constraints. *Int. J. Comput. Vis.* 95, 74–85 (2011)
9. Scaramuzza, D., Siegwart, R.: Appearance-Guided Monocular Omnidirectional Visual Odometry for Outdoor Ground Vehicles. *IEEE Transactions on Robotics* 24, 1015–1026 (2008)
10. Tardif, J.P., Pavlidis, Y., Daniilidis, K.: Monocular Visual Odometry in Urban Environments Using an Omnidirectional Camera. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008)*, pp. 2531–2538 (2008)
11. Caron, F., Duflos, E., Pomorski, D., Vanheeghe, P.: GPS/IMU Data Fusion Using Multi-sensor Kalman Filtering: Introduction of Contextual Aspects. *Information Fusion* 7, 221–230 (2006)
12. Kim, S., Yoon, K., Lee, D., Lee, M.: The Localization of a Mobile Robot Using a Pseudo-lite Ultrasonic System and a Dead Reckoning Integrated System. *International Journal of Control, Automation and Systems* 9, 339–347 (2011)
13. Wei, L., Cappelle, C., Ruichek, Y., Zann, F.: GPS and Stereovision-Based Visual Odometry: Application to Urban Scene Mapping and Intelligent Vehicle Localization. *International Journal of Vehicular Technology* 17 (2011)
14. Suzuki, T., Kitamura, M., Amano, Y., Hashizume, T.: 6-DOF Localization for A Mobile Robot Using Outdoor 3D Voxel Maps. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pp. 5737–5743 (2010)
15. Yekeun, J., Yunsu, B., Jun-Sik, K., In-So, K.: Complementation of Cameras and Lasers for accurate 6D SLAM: From Correspondences to Bundle Adjustment. In: *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3581–3588 (2011)
16. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* 60, 91–110 (2004)
17. Mei, C., Rives, P.: Single View Point Omnidirectional Camera Calibration from Planar Grids. In: *IEEE International Conference on Robotics and Automation*, pp. 3945–3950 (2007)
18. Siegwart, R., Nourbakhsh, I.R.: *Introduction to Autonomous Mobile Robots*. Bradford Company (2004)
19. Labrosse, F.: The visual compass: Performance and Limitations of an Appearance-Based Method. *Journal of Field Robotics* 23, 913–941 (2006)
20. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and Chamfer Matching: Two New Techniques for Image Matching. In: *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 659–663. Morgan Kaufmann Publishers Inc., Cambridge (1977)
21. Hoang, V.-D., Vavilin, A., Jo, K.-H.: Fast Human Detection Based on Parallelogram Haar-Like Feature. In: *The 38th Annual Conference of the IEEE Industrial Electronics Society* (2012)
22. Solà, J., Vidal-Calleja, T., Civera, J., Montiel, J.: Impact of Landmark Parametrization on Monocular EKF-SLAM with Points and Lines. *Int. J. Comput. Vis.* 97, 339–368 (2012)
23. Thrun, S., Burgard, W., Fox, D.: *Probabilistic robotics*. MIT Press, Cambridge (2005)
24. Civera, J., Grasa, O.G., Davison, A.J., Montiel, J.M.M.: 1-Point RANSAC for Extended Kalman Filtering: Application to Real-Time Structure from Motion and Visual Odometry. *J. Field Robot.* 27, 609–631 (2010)

Automatic Reconstruction of Two-Dimensional Broken Objects

Yanjuan Zhu

School of Aerospace Engineering and Applied Mechanics,
Tongji University, Shanghai 200092, China
zhu_yj@tongji.edu.cn

Abstract. This paper proposes a novel approach to automatic reconstruction of two-dimensional fragments. The contour matching task in fragment reassembly is reduced into a feature segment matching problem. Efficient techniques have been applied to find the proper matching segments. Specifically, the extracted contour of every fragment is defined as a set of feature segments on it. The probability of matching for fragments is measured by the Hausdorff distance between the corresponding feature segments. If the two feature segments are matchable, one of them is rotated and translated to be coincided with another, and then the overlap contact is detected automatically under this transformation. Fragments without overlap is merged to form a new one through contour updating. Then the matching process is continued until the original object is reconstructed. Experimental results demonstrate the algorithm is reliable and efficient.

Keywords: Contour extraction, Convolution integral, Relative curvature, Contour matching, Feature segment.

1 Introduction

Reconstruction of unknown objects from a large number of irregular fragments is a tedious and laborious task, when it is finished merely by hand. With the computer's help, the automatic reassembly of fragmented objects would alleviate the manual effort evidently. The related technology has important applications, including failure analysis of aero-craft, palaeontology research, synthetic drug design and geology study. It also evaluates the practical and theoretical worth of this technology.

A useful general-purpose method for fragment reassembly should make accurate and reliable reconstruction of original object.

When two fragments match within the given tolerance, there exists a transformation consisting of rotation and translation, which will cause one of the fragments to match the other and form a new fragment. ①The contour of merged fragment should be updated in real-time. ②The method should be insensitive to the reconstruction sequence. ③The method should be efficient.

We propose a new method for automatic reconstruction of broken objects in two dimensions. This technique is based on using feature segments to represent fragment contour. The fragment reassembly proper is measured by the Hausdorff distance between the feature segments. For feature segments matching within the given tolerance, we compute the transformation that maps one feature segment to the other and makes the feature segments coincide. When the two fragments have no overlap contact under the transformation, it verifies the two fragments are of a good match. Therefore, they are merged and form a new piece. Otherwise the fragments are an invalid matching, though they are satisfied with the tolerance request. This process is repeated until the fulfillment of reconstruction.

1.1 Related Work

The problem of two-dimensional fragment reassembly can be regarded as the generalization of jigsaw puzzles. As a cute exercise, the computer solution of jigsaw puzzles goes back at least to 60's in the last century [1], and due to the rapid development in computer science, it receives extensive study in robotics and machine vision. However, the techniques used are applicable only to relatively regular shapes of jigsaw puzzles, such as smooth edges and well-defined corners. In particular, Burdea and Wolfson[2] subdivide the boundary curve of each piece into four subcurves and then search for the counterpart for every subcurve. Wenster et.al[3] justify the two pieces are matched by utilizing isthmus as a global feature. David Goldberg et.al[4] measure the probability of matching through center point, inflection point and tangent point on each piece. But Fragments as mural paintings, which broken process is random, the problem of assembly tasks is particularly evident, where two objects match along a common boundary without any obvious start point and end point for the common subcurve. Therefore, it is clear that these techniques can not be directly used for fragments reassembly.

Weixin Kong et.al[5] and H.C.G. Leitao et.al[6]deal with the problem of two-dimensional fragment reassembly respectively. The former compute the affinity of two pieces but don't manage the overlap detection. The actual operation has to be done by human interaction. The latter is focus on the first step which is the identification of matching pairs, so the overlap constraint is not considered. Nevertheless, the overlap detection process is unavoidable in the consequent reassembly of fragments.

2 Fragment Representation

Establishment of the description for fragment in computers is of central importance in the process of matching. It should be suitable for computer to handle, and meanwhile satisfy the requirements relating to describing shape and measuring the matching level.

2.1 Contour Extraction

In order to get a digital representation of the fragments, each of them are scanned by a standard flatbed scanner in an arbitrary orientation. The boundary of each piece is extracted from the corresponding image. On the basis of Canny edge detector [7], a geometric method is developed to obtain the contour of fragment accurately. Canny approach is one of the most widely used edge detection. It can find the absolute majority of object contour in an image, but it inevitably misses some important and obvious edges during boundary detection [8], which cause gaps in the object contour.

Suppose G is a two-dimensional Gaussian[9].

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

G_n is a first derivative of G in the direction n .

$$G_n = \frac{\partial G}{\partial n} = n \cdot \nabla G = n \cdot \begin{bmatrix} \partial G/\partial x \\ \partial G/\partial y \end{bmatrix} \tag{2}$$

The direction n can be estimated in formula (3)

$$n = \frac{\nabla(G * f)}{|\nabla(G * f)|} \tag{3}$$

Where f is the image function. The edge location is at the local maximum of the image f convolved with the operator G_n .

$$\frac{\partial}{\partial n} G_n * f = 0 \tag{4}$$

Namely

$$\frac{\partial^2}{\partial n^2} G * f = 0 \tag{5}$$

However, if gradient magnitude of edges is the local maximum in the image, while it is not in the gradient direction, it will not be picked up accordingly. Gaps on the contour result in curvature calculation are rather difficult, and hence affect the quality of contour matching or fragment reassembly.

The first step in the geometric method is to assign specific coordinates for every edge point detected by Canny operator. An arbitrary point P_i on the contour is selected as starting point and its coordinates are registered. The point P_{i+1} with the smallest distance to P_i is found and the coordinates are calculated according to the distance. Then P_{i+1} is regarded as the current starting point. The closest point P_{i+2} in its neighborhood is discovered and the coordinates are also calculated. The above process is repeated until all the all the edge points have corresponding coordinates. All of the edge points are stored in an array.

For the relationship between each pair of adjacent points, the following three categories should be considered further. ① The x -coordinate is the same, whereas the y -coordinate differs. The difference of y -coordinates is calculated. ② The y -coordinate is the same, whereas the x -coordinate differs. The difference of x -coordinates is calculated. ③ Both of the x and y coordinates are different. The sum of the two differences is calculated.

For condition ① or ② to take place, if the difference is more than \mathcal{E}_1 , it shows there exists a gap between the two points. For condition ③, if the difference is more than \mathcal{E}_2 , it also shows that there is a gap. The value of \mathcal{E}_1 and \mathcal{E}_2 are dependent on the space between pixels. All these gaps are filled according to the orientation of the original contour. The points that represent a two-dimensional contour are naturally ordered as a sequence, thus enabling application of efficient curvature calculation algorithms, which exploit such an order.

2.2 Curvature Calculation of Discrete Point

The curvature of a curve at any point is defined as the rate of change in tangent vector direction with respect to arc length [10]. It measures how much it is necessary to turn the steering wheel when driving along the curve. For two-dimensional contour, relative curvature can describe it more exactly. The curvature may be positive or negative, depending on whether the curve is bending toward the normal vector or away from the normal vector.

For the purpose of calculating the curvature of every point on the fragment contour, we present a convolution approach, which utilizes an appropriate linear interpolation to resample the contour to perform the curvature calculation. The size of convolution window is set firstly. Its center is the point which curvature is being computed. Convolution integral is the process of weighted averaging for the points within the window.

Suppose a discrete contour

$$\vec{r} = (x(u), y(u)) \quad u \in (0, 1) \tag{6}$$

Where u is an arclength parameter. Let $x(u)$, $y(u)$ convolve with Gaussian kernel respectively.

$$\vec{r}_\sigma = (X(u, \sigma), Y(u, \sigma)) \tag{7}$$

Where $X(u, \sigma) = x(u) * G(u, \sigma)$, $Y(u, \sigma) = y(u) * G(u, \sigma)$

Then curvature of the curve \vec{r}_σ is [11]

$$K(u, \sigma) = \frac{X_u(u, \sigma)Y_{uu}(u, \sigma) - X_{uu}(u, \sigma)Y_u(u, \sigma)}{(X_u(u, \sigma)^2 + Y_u(u, \sigma)^2)^{3/2}} \tag{8}$$

A linear interpolation method is put forward to resample the contour in specifying the computation. Given two points r_i and r_{i+1} the inserted point r_t is decided by Eq.(9).

$$r_t = \left(1 - \frac{s - \sum_{i=0}^j D[i]}{\sum_{i=0}^{j+1} D[i] - \sum_{i=0}^j D[i]}\right)r_i + \frac{s - \sum_{i=0}^j D[i]}{\sum_{i=0}^{j+1} D[i] - \sum_{i=0}^j D[i]}r_{i+1} \tag{9}$$

Where s is the interval between inserted point and the center point, j is the function of s . The value of $D[i]$ is the distance between two adjacent points. All of the pointwise curvature can be obtained, as the window is slid along the contour.

The proposed method is utilized to compute the minimal, maximal and average value of the curvature at the one thousand points, which is randomly distributed on the circumference whose radius is 8 centimeter. The result is

$$k_{\min} = 0.123932, \quad k_{\max} = 0.130708, \quad k_{\text{avg}} = 0.125990 \quad (10)$$

It indicates that the accuracy of the curvature estimate can be directly used for reliable matching.

2.3 Feature Detection and Contour Segmentation

We present a feature point detection procedure based on the local extremum of the relative curvature. These are the points where the curve is changing at its highest angular speed. This procedure is implemented in two stages for avoiding the number of feature point is a little larger and its distribution is dense with respect to the contour perimeter. All of the local extremum points are detected in the first stage. When any two adjacent local extremum points have the same sign in relative curvature, the one that the absolute value is smaller is removed. The second stage is to consider the absolute value of relative curvature for these remainder points. When the curvature at this point is above a set threshold and also above the twice of the minimum curvatures at its two neighbor points, it will be selected as feature point.

Feature point and some other points in its neighborhood constitute a feature segment. According to the locations of feature points on the contour, the feature segments are classified into two categories. ① Full feature segment: The intersection of two fractures forms a feature point. The corresponding feature segment consists of the feature point and some other points at its both side edges. ② Semi-feature segment: The intersection of a fracture and the original object contour forms a feature point. The feature segment only consists of the feature point and some other points at fracture side.

This representation by feature segments is extremely compact as compared to the full two-dimensional specification, which might otherwise be required to represent the real-world object.

3 Fragment Reassembly

We use the Hausdorff distance to measure the matching probability of feature segments belonging to different fragments. At ideal condition, each pair of matchable segments is almost the same. But some small fragments may be lost during the broken process, and due to imperfections in the digitization environment. Therefore, the Hausdorff distance between any two matchable segments is less than the given tolerance ε which is non-zero.

3.1 The Hausdorff Distance

The Hausdorff distance [12] is a max-min distance between two arbitrary points sets. It is a similarity measure of them. Given two bounded sets $A=\{a_1, a_2, \dots, a_p\}$ and $B=\{b_1, b_2, \dots, b_q\}$, the Hausdorff distance is taken as

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{11}$$

Where the function $h(A, B)$ is called the directed Hausdorff distance from A to B. $h(B, A)$ is similar.

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{12}$$

$$h(B, A) = \max_{b \in B} \min_{a \in A} \|b - a\| \tag{13}$$

Here, we assume that the underlying metric $\|\bullet\|$ is Euclidean metric, but the consideration is also valid for other commonly used metrics. This shows it assigns to each point of one set the distance to its closest point in the other and takes the maximum of all these values. Thus the Hausdorff distance measures the degree of match between two sets.

3.2 Measuring the Probability of Matching

Two pieces, *Piece1* and *Piece2*, are selected for example purposes. The two contours are defined as $Piece1 = \{m_1, m_2, \dots, m_p\}$ and $Piece2 = \{n_1, n_2, \dots, n_q\}$ respectively, where $m_i (i = 1 \dots p)$ and $n_j (j = 1 \dots q)$ are feature segments. Moreover, $m_i = \{k^i_1, k^i_2, \dots, k^i_c \dots k^i_l\}$ and $n_j = \{K^j_1, K^j_2, \dots, K^j_c \dots K^j_l\}$ represent the curvature of every point on the *i*th, the *j*th feature segments respectively. l denotes the length of feature segment. k^i_c and K^j_c are curvatures of center points. Obviously, for semi-feature segment, C is equal to l . The directed Hausdorff distance is given by

$$h(m_i, n_j) = \max_{k^i_s \in Piece1} \min_{K^j_t \in Piece2} |k^i_s - K^j_t| \quad s, t \in 1 \dots l \tag{14}$$

$$h(n_j, m_i) = \max_{K^j_s \in Piece2} \min_{k^i_t \in Piece1} |K^j_s - k^i_t| \quad s, t \in 1 \dots l \tag{15}$$

In general, $h(m_i, n_j)$ is not the same as $h(n_j, m_i)$. For the structure of the two sets is consistent, they are equivalent. Hence, the Hausdorff distance is

$$H(m_i, n_j) = \max(h(m_i, n_j), h(n_j, m_i)) = \max_{k^i_s \in m_i} \min_{K^j_t \in n_j} |k^i_s - K^j_t| \quad s, t \in 1 \dots l \tag{16}$$

If $H(m_i, n_j) \leq \varepsilon$, it means the two feature segments are matching within the tolerance. There is a problem of corresponding point search. For every feature point detected on one contour, its corresponding point with the matching contour may not be a feature point. The reason lies in the fact that the criterion of feature point involves the curvature and contour trend. We transform this problem to a semi-feature segment problem. In other words, the corresponding point, which has the same value of curvature and opposite direction with respect to that of the feature point, is searched at first. Then the curvature of every point in one sequence is compared to every curvature in the other sequence. If only half of them are similar, it is inferred that the fragments can be reassembled along the partial contour.

3.3 Contour Updating and Sequence Matching

Euclidean transformation is applied to reassemble the matching fragments, and the contour of the glued fragment is updated accordingly. However, not all of the feature segments matching within the tolerance ϵ are actual matching. Usually, there are some feature segments with similar curvatures between different fragments. Nevertheless, they could not be really reassembled. An approach to detect overlap contact is developed to retain valid matchings and rule out invalid ones.

Suppose m_i of *Piece1* and n_j of *Piece2* are a potential pair. The detail of this approach is illustrated below. Step1. Produces the rotation and translation parameters needed to match n_j to m_i . Step2. Fulfill the Euclidean transformation. Step3. *Piece2* is represented by a polygon consisting of all the feature points on the contour. Step4. Draw a line through the vertex of feature polygon which is parallel to the x-axis. Calculate the intersection of the line and the boundary curve of *Piece1*. Add up the number of intersections. Step5. If the feature point lies on a line between the even point and odd point, it means the two fragments have no overlap after reassembly. It will then continue the matching process with the updated contour. On the other hand, if the feature point of *Piece2* lies on a line between the odd point and even point, it means the two fragments have overlap under the current transformation.

When all of the potential pairs result in overlap contact, it further proves that the two pieces are not actual matching.

4 Experimental Results

A subsequent series of experiments tested and evaluated the performance of the matching algorithm with real two-dimensional contour data. Fig.1 shows the images of several pieces taken from a map of China and the contours extracted from these pieces which have been improved by the geometric method proposed in this paper.

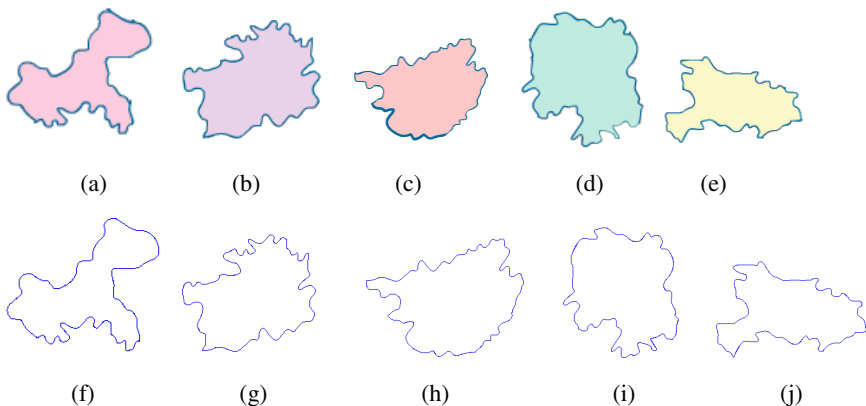


Fig. 1. (a)~(e)Five images of puzzle pieces taken from Chinese map. (f) ~ (j)Contour curves extracted from the puzzle pieces.

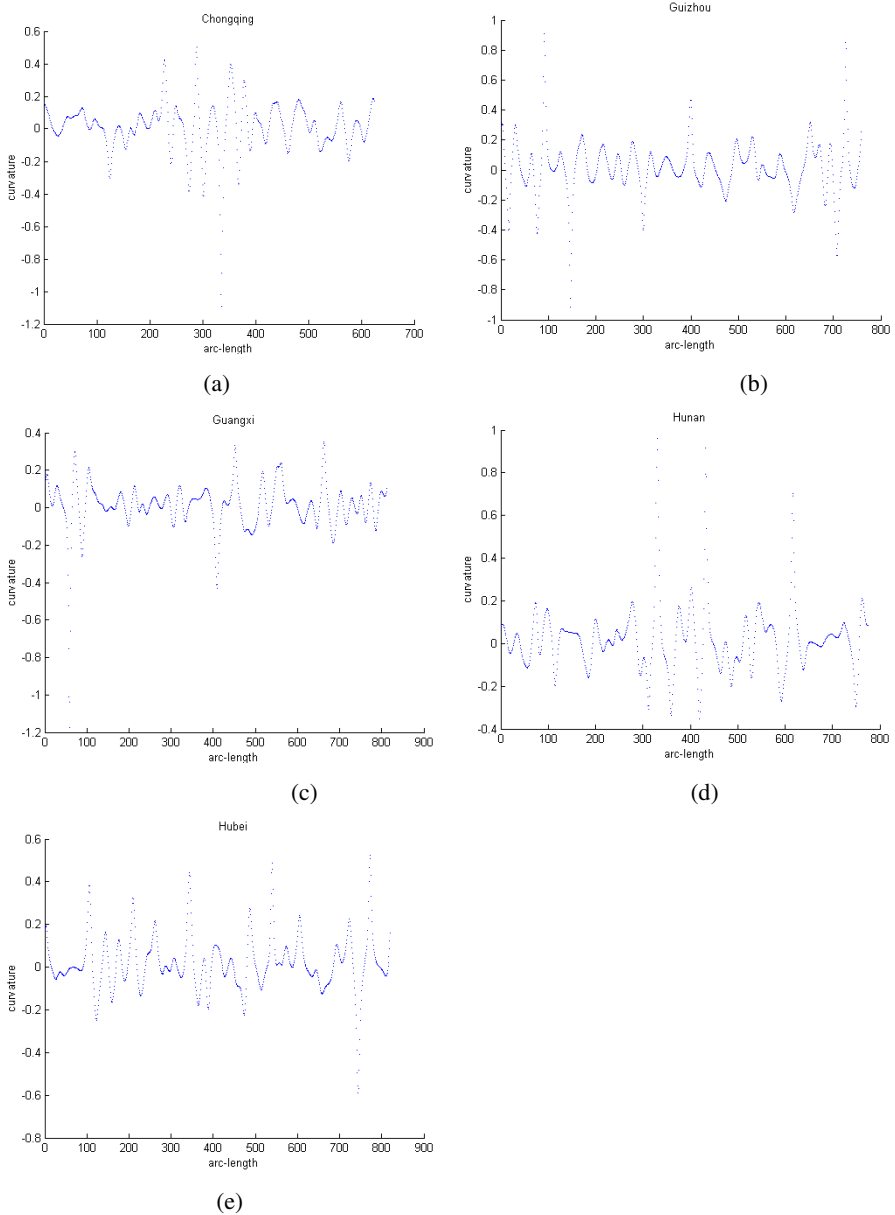


Fig. 2. (a)~(e) Curvature computed from the extracted contours

Fig.2. shows the curvature derived from the contour for the five different pieces. The contours of the pieces are matched against each other in order to determine how these pieces should be assembled together. Fig.4~7 show the results of matching individual pieces and finding the best match. The red points are feature points of the contour.

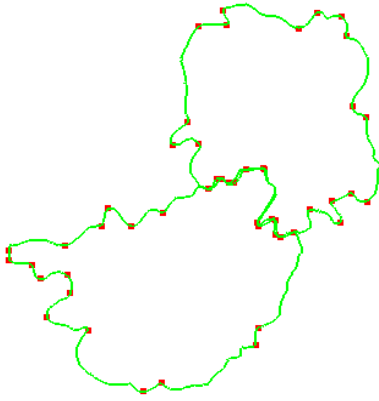


Fig. 3. The result of matching the two contours

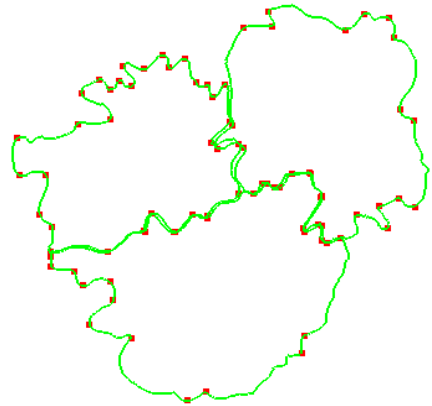


Fig. 4. The result of matching the three contours

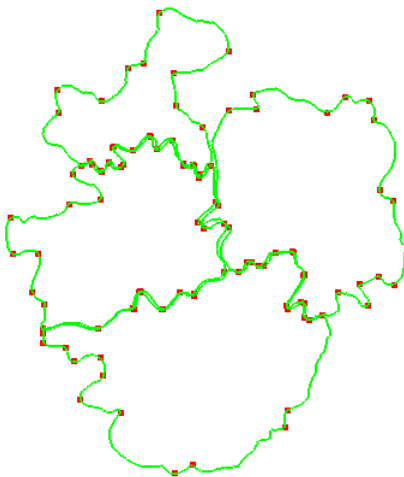


Fig. 5. The result of matching the four contours

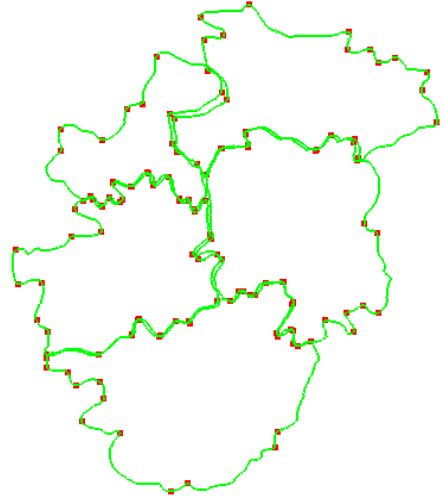


Fig. 6. The result of matching the five contours

5 Conclusion

We have proposed an algorithm for solving the practical problem of fragment reassembly. Feature segments information is used to represent fragments and this representation is invariant to translation, rotation and scaling. The Hausdorff distance is utilized to measure the matching probability of feature segments from different fragments. In contrast with previous work, this algorithm not only enables efficient

contour comparison and matching, but also identifies the overlap contact with respect to fragments under Euclidean transformation, which is a key step in the full reconstruction of original object. As demonstrated, our technique finds accurately the matches between the fragments and fulfills the global consistency between different matches. In addition, this method can be applied to the problem of matching one contour against another (two-contour matching), as well as matching an observed contour against a large data base of model contours.

References

1. Freeman, H., Gardner, L.: Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *IEEE Trans. on Electronic Computers* 13, 118–127 (1964)
2. Burdea, G.C., Wolfson, H.J.: Solving Jigsaw Puzzles by a Robot. *IEEE Trans. on Robotics and Automation* 5, 752–764 (1989)
3. Webster, R.W., LaFollette, P.S., Stafford, R.L.: Isthmus critical points for solving jigsaw puzzles in computer vision. *IEEE Transactions on Systems Man and Cybernetics* 21, 1271–1278 (1991)
4. Goldberg, D., Malon, C.: A global approach to automatic solution of jigsaw puzzles. In: *Symposium on Computational Geometry*, pp. 82–87 (2002)
5. Kong, W.X., Kimia, B.B.: On Solving 2D and 3D Puzzles using Curve Matching. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii (December 2001)
6. Leitão, H.C.G., Stolfi, J.: A multi-scale method for the reassembly of two-dimensional fragmented objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(9), 1239–1251 (2002)
7. Canny, J.F.: A Computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
8. Ding, L.J., Goshtasby, A.: On the Canny edge detector. *Pattern Recognition* 34(3), 721–725 (2001)
9. Pan, W.J.: *Fourier Analysis and Application*. Peking University Press (May 2000)
10. Shi, F.Z.: *Computer Aided Geometric Design & Non-uniform Rational B-spline*, pp. 17–25. Beijing University of Aeronautics and Astronautics Press, Beijing (1994)
11. Mokhtarian, F., Mackworth, A.K.: A Theory of Multiscale, Curvature-Based Shape representation for Planar Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(8), 789–805 (1992)
12. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(9), 850–863 (1993)

An Efficient Natural Image Deblurring Algorithm

Rajesh R. Pillai¹, Vandana Dixit Kaushik², and Phalguni Gupta¹

¹ Department of Computer Science & Engineering,
Indian Institute of Technology Kanpur, Kanpur 208 016, India

² Department of Computer Science & Engineering,
Harcourt Butler Technological Institute, Kanpur 208002, India
vandanadixitk@yahoo.com, pg@iitk.ac.in

Abstract. This paper has considered the problem of deblurring of an image which is an ill-posed and challenging problem due to not only the large number of unknowns but also non-availability of more number of images of the same scene or objects. It uses the Variational Bayesian approach to optimize the posterior probability and to derive the most probable Point Spread Function (PSF). Lucy Richardson algorithm [1] has been modified to get the deblurred image. The algorithm is found to be very effective for natural images.

Keywords: Deblurring, Point Spread Function, Variational Bayesian, Signal to Noise Ratio.

1 Introduction

Relative motion between camera and the subject is a primary reason for blurring in an image. It can take convoluted paths and thus making the restoration more complex. Restoring of blurred image involves two components, namely, identification of blur and restoration of image using the blurring parameters. To restore a blurred image successfully, blurring function, referred to as Point Spread Function (PSF), needs to be estimated accurately. PSF is the response of an imaging system to a point source or it can be said as the impulse response of a focused optical system. It is non-parametric and spatially varying.

Deblurring is an ill-posed problem because of the number of unknown parameters is more than the available parameters. Attempts have been made to address the problem of blind deconvolution for deblurring of a natural image. Recent algorithms have achieved dramatic progress. However, there exist many aspects of the problem which are still challenging and hard to solve. Lokhande et. al. [2] have worked on identification of motion blur parameters using frequency domain analysis and have tried to reconstruct the PSF using the length and angle information. It may fail to perform well for natural images because the algorithm assumes PSF to be perfectly box (linear) which is not the case in many natural images. The algorithm also does not cater for varying noise levels. Joshi et.al. [3] have tried to estimate the PSF using

sharp edge prediction. They have tried to predict the ideal edge by finding the local maximum and minimum pixel intensities. It does not perform well for larger blurs. Levin et.al. [4] have proposed an algorithm to deblur a blurred image using image statistics. They have shown that the direction of motion blur is the direction with minimal derivative variation and the value which gives the maximum likelihood of the derivatives is the blur length. It gives good results only for box kernels which are the characteristics of perfect motion blurs. Blurs may not be always motion blurs and most of the motion blurs do not have perfect box PSF. Fergus et. al.[5] have approached the problem using a Variational Bayesian approach for PSF estimation. Shan et. al.[6] have used a semi maximum a-posteriori (MAP) algorithm which is used to get a point estimate of the unknown quantity based on empirical data. They have used a Gaussian prior for natural image and edge reweighting and iterative likelihood update for approximation of latent image. It does not perform well for all images which are sparse or a bit away from Gaussian. Yuan et.al. [7] have used two sets of images (one blurry and one noisy) to recover the original image. A comprehensive literature review to approach a deconvolution problem can be found in [8]. Miskin and Mackay [9] have used ensemble learning algorithm to extract hidden images from a given image. They have used Variational Bayesian approach to do ensemble learning.

2 General Blur Model

There are two types of deblurring approaches namely, non-blind deblurring and blind deblurring. In non-blind deblurring, knowledge of the Point Spread function (PSF) is available which is not the case with blurred deblurring. Let F , H and N be the true image, the point spread function (PSF) and the additive random noise respectively. Then the measured image G is given by

$$G = F * H + N \quad (1)$$

That means, for each pixel (x,y) , the intensity value of the measured image G can be computed by

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

$$= \sum_{m=0}^{m-1} \sum_{n=0}^{n-1} f(x, y)h(x - m, y - n) + n(x, y) \quad (3)$$

where $*$ is convolution operation.

If one considers frequency domain, then one gets

$$G(u, v) = F(u, v)H(u, v) + N(u, v) \quad (4)$$

where parameters on the right hand side of Equation (4) are unknown. For a perfect motion blur, parameters to be considered are the length and the direction of blur. In this case if it is assumed that there is no noise, it is easy to reconstruct the PSF using the length and the angle parameters and to do a non-blind deblur operation. Some of the typical PSFs are shown in Figure 1. There are four different types of Point Spread Functions, namely Motion Blur, out of Focus Blur, Gaussian Blur and Scatter Blur that are very basic and generic in nature. Out of those four, two functions belong to the category of sharp edged PSFs and remaining two belong to the category of smooth edged PSFs. These PSF models are in time domain.

3 Proposed Algorithm

Lokhande et. al [2] have used frequency domain to estimate the blurring parameters such as length and angle. It uses the spectrum of an image to analyze the blur and considers hough transform to determine the blurring angle. But spectrum of an image does not remove all noise and displays lines which do not represent the blur direction in the image. Further, hough transform does not perform well in such a case and is computationally intensive. If we use the cepstrum of the image instead of spectrum and consider radon transforms instead of hough transforms, computation becomes easy and we get better estimates of blurring angle. The binary cepstrum image is rotated by the estimated angle and average of each column is taken. The distance between the zero-crossings represents the inverse of the length parameter. The algorithm is as follows:

Algorithm Blurring-Parameters

Convert the image to grayscale

Calculate Log and square of the image

Calculate Inverse Fast Fourier Transform to get the cepstrum.

Convert to binary

Apply radon transform for various angles

Find the angle at which the radon transform value is maximum to get blurring angle

Calculate average along each column

Find the distance between the zero crossings to get the periodicity to get blurring length

Using parameters like blurring length and angle, the PSF can be constructed and the latent image can be reconstructed using any well known deconvolution algorithm such as Wiener Filter which is computationally less expensive. Approximating the correct value of Noise to Signal Ratio dictates the quality of the output of the Wiener Filter. We have approximated Noise to Signal Ratio (NSR) as follows:-

$$\text{NSR} = \frac{1}{20 \log_{10} \frac{A-B}{\sigma}} \quad (5)$$

where A , B and σ are the maximum, the minimum and the standard deviation of the intensities of the image.

This algorithm works well for synthetically generated blurs but fails for natural blurs. The main reason for this is the fact that most of the natural blurs are not perfect motion blurs which have an angle and length. These blurs are either out of focus blurs or non-linear motion blurs due to camera shake, or random blurs. The biggest challenge is to know the type of blur and then decide the method of finding the PSF which has random shape.

This section proposes an efficient blind deconvolution algorithm which can be used to deblur any natural image without the knowledge of the PSF. It consists of 3 major steps and they are (i) selection of region of interest, (ii) PSF estimation, and (iii) generation of deblurred image.

3.1 Region of Interest Selection

Any blurred image I can be either colour or grayscale. In order to reduce the computational cost, the image I is converted to grayscale. However, the colour information is retained so that the same can be regenerated to obtain the deblurred coloured image. An inverse gamma correction is done to the image during pre-processing. We use a gamma value of 2.2, which is decided empirically.

Since processing of the whole image is computationally intensive, only a specific patch of the image is considered for deblurring. Let the gray scale blurred patch of the image I be G . One should be careful while selecting the patch region so that it does not contain any saturated region because this type of area does not any gradient variation. Gradients of an image are considered to represent better change of patterns than image intensities. Gradients in horizontal and vertical directions are calculated and concatenated to get the gradient matrix inverted *triangle* G .

$$\nabla G = [\nabla G_x, \nabla G_y] \quad (6)$$

One can use simple gradient kernel $[1 \ -1]$ and $[-1 \ 1]$ for horizontal and vertical gradients respectively.

3.2 PSF Estimation

The Point Spread Function (PSF) (H) is initialized to a 3X3 matrix which is selected according to the selection of vertical or horizontal directions. The value of H for horizontal direction is

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and for vertical direction is

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

The extent of processing depends on the size of the PSF which is initialized to an approximate value ϕ . The estimation of PSF is done on a coarse to fine manner by varying the image resolution (at the coarsest scale H is 3×3). The number of iterations is decided by the scale S which is given by

$$S = -2 \log_2 \left(\frac{3}{\phi} \right) \quad (7)$$

where ϕ is the approximate size of the PSF. The initial estimate of the latent image is obtained by variational bayesian inference algorithm [9] having fixed H . The inference is done at each scale to find out converged values of H and S . These converged values act as the initial values of the next inference at next scale. At the finest scale, the full resolution Kernel is achieved. Minor inaccuracies or noise in the estimated PSF can affect the output. The algorithm uses a threshold for the PSF so that it is refined and noisy values are suppressed to zero.

$$Threshold = \frac{C}{15} \quad (8)$$

where C is the maximum intensity value of H .

3.3 Recovering Latent Image

A smoothing function is applied to the sharp edges of the image to reduce sharp the ringing artifacts. The non-blind accelerated damped Lucy-Richardson Algorithm is used to deblur the whole image using the estimated Point Spread Function (PSF). The number of iterations to be used in the Lucy Richardson algorithm [1] depends on the extent of blur. Higher the extent of blur, larger is the number of iterations. A default value of 10 has been used in this algorithm. Gamma correction is applied to the image to revert the inverse-gamma applied earlier. Histogram equalisation of the recovered image is done to match with the original image. The recovered image is a colour image in case the input image is coloured. The algorithm is given below.

*Algorithm Deblurring**Convert image to grayscale**Perform inverse gamma correction to image, $\gamma=2.2$* *Select ROI**Compute horizontal gradients $[\nabla G]_x$ and vertical gradients $[\nabla G]_y$* *Concatenate gradients $[[\nabla G]_x [\nabla G]_y]$* *Compute number of scales S* *Loop over scales starting with coarsest to fine**Rescale gradients**Compute initial estimate of latent image F using variational bayesian inference holding H fixed.**Rescale estimated F and H from previous scale**Modify latent image and PSF using variational bayesian inference**Threshold PSF**Edge smoothening of G to reduce ringing artifact**Apply accelerated damped Lucy Richardson Algorithm to deblur**Perform gamma correction to the recovered image, $\gamma=2.2$* *Do histogram equalisation to get the original colour values.***4 Experimental Results**

To test the performance of the proposed algorithm, we have considered 7 blurred images and they are Car1, Car 2, Bike, Gate, Peacock, Ground and Face. To verify the correctness of the proposed algorithm, we have synthetically blurred the images. Results are shown in Figure 1 - Figure 4. Although the quality of the image as whole has not improved much, it can be clearly seen that information like the number plate of the vehicle, name written on the gate, the title written on the erected board and the peacock's feathers are available of the recovered image.



Fig. 1. Blurred Images of Number Plates with their Respective Deblurred Images

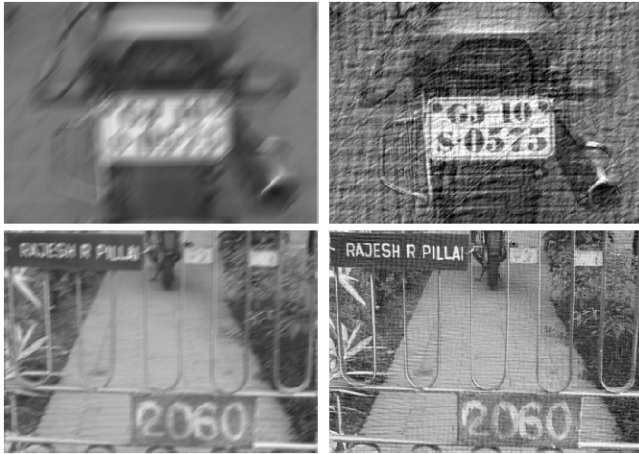


Fig. 2. Blurred Images of Number Plates and Gate with their Respective Deblurred Images



Fig. 3. Blurred Images of Ground, Face with their Respective Deblurred Images



Fig. 4. Blurred Image of Peacock with its Deblurred Image

5 Conclusion

This paper has proposed an efficient algorithm to deblur a natural image. However it has made the assumption of uniform blur throughout the image. It has used the Variational Bayesian approach to find the most probable Point Spread Function. A modified Lucy Richardson algorithm is used to perform the deconvolution to get the deblurred image. It has been tested on several sets of blurred natural images and found that the algorithm could deblur the images effectively.

References

- [1] William, H.R.: Bayesian-Based Iterative Method of Image Restoration. *Journal of Optical Society of America* 62(1), 55–59 (1972)
- [2] Lokhande, R., Arya, K.V., Gupta, P.: Identification of Parameters and Restoration of Motion Blurred Images. In: *Proceedings of the 2006 ACM Symposium on Applied Computing*, pp. 301–305 (2006)
- [3] Joshi, N., Szeliski, R., Kriegman, D.J.: PSF Estimation Using Sharp Edge Prediction. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8 (2008)
- [4] Anat, L.: Blind Motion Deblurring Using Image Statistics. In: *Proceedings of Advances in Neural Information Processing Systems, NIPS (2006)*
- [5] Rob, F., Barun, S., Aaron, H., Sam, T.R., William, T.F.: Removing Camera Shake from a Single Photograph. *Proceedings of ACM SIGGRAPH 2006* 25(3), 787–794 (2006)
- [6] Shan, Q., Jia, J.Y., Aseem, A.: High-quality Motion Deblurring from A Single Image. *Proceedings of ACM SIGGRAPH 2008* 27(3), 1–10 (2008)
- [7] Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image Deblurring with Blurred/Noisy Image Pairs. *Proceedings of ACM SIGGRAPH 2007* 26(3) (2007)
- [8] Kundur, D., Hatzinakos, D.: Blind Image Deconvolution. *IEEE Signal Processing Magazine* 13(3), 43–64 (1996)
- [9] James, M., David, J.C.: Ensemble Learning for Blind Image Separation and Deconvolution. In: Girolani, M. (ed.) *Independent Component Analysis*. Springer (2000)

A Heuristic Technique for Performance Improvement of Fingerprint Based Integrated Biometric System

Kamlesh Tiwari, Soumya Mandi, and Phalguni Gupta

Department of Computer Science and Engineering, Indian Institute of Technology,
Kanpur, Kanpur 208016, India
{ktiwari, smymandi, pg}@cse.iitk.ac.in

Abstract. Integrated biometric system coherently connects multiple independent biometric systems to improve upon security and usability. But due to some involved factors the performance of overall system get degraded. This paper proposes a heuristic technique to improves the performance of fingerprint based integrated biometric system. One of the main advantage of the proposed technique is that it does not ask to modify the underline identification approach of the independent systems. The technique has been tested by constructing an integrated system with three commercial available independent biometric systems and the database used is IITK-Rural containing fingerprint images of 500 subjects collected over two sessions on three scanners. The proposed heuristics improves the equal error rate (EER) of integrated system from 21.78% to 15.67%.

Keywords: Biometrics, Heuristics, Integrated systems, Performance, ROC Curve.

1 Introduction

Biometrics offers an intuitive way of human identification with the help of his behavioral or physiological characteristics. It has many applications in access control and personal authentication. Under typical settings, two independent biometric systems such as one for banking and another for marking office attendance work in heterogeneous mode and they do not share any data. But, there is a need of sharing and system integration to improve interoperability, security, privacy and usability of the system. Sharing enables the development of integrated systems where all independent systems coherently collaborate and offer transparent services. In integrated system environment, user can register his biometric sample at any participating independent system and later the same sample can be used by all available systems as gallery. Under this setting, a person might get enrolled at one independent system and get verified by another independent system. It involves communication to take place between different modules of different systems. Advantage of integrated system is the possible low duplication of data which increases system entropy and thereby improves security and privacy of the system. It also improves the usability of the system because of increased number of access points. One of the challenges faced during integration is the

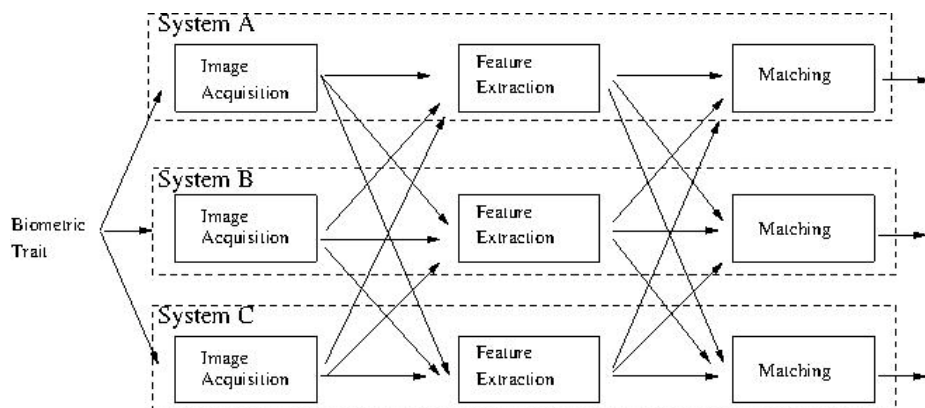


Fig. 1. Integrated system involving three independent systems A, B and C

negative impact on overall system performance. This is primarily because of the involvement of heterogeneous scanners, different feature extractors and matchers in the system. Technology to capture the biometric data varies with sensor. For example, there are sensors which are based on optical imaging and there are sensors based on solid-state technology. Modeling of fingerprint feature [1] and matching strategy [2,3] can also vary. It has been observed experimentally that feature extraction and matching routines of different systems are faithful to their scanners. It has been shown in [4] that under certain assumptions, change in sensor does not have an impact on the performance of a matching system.

The problem statement addressed in this paper is to devise a technique that improves the overall system performance of integrated system without modifying underline technology of independent systems. This paper proposes a heuristic based normalization strategy which bounds the error by an upper limit, thereby improving the performance of the integrated system. It involves registration of new independent biometric system with the already existing integrated system. The effect of the strategy is studied using IITK-Rural database where the images have been captured with three independent systems. Rest of the paper is organized as follows. Section 3 explains the proposed technique in detail, whereas the next section analyses the experimental setup and results. Conclusions are presented at the last section.

2 Mathematical Background

Any biometric system like face [5-7], iris [8,9], fingerprint [10,11], palm [14,15], knuckle [14,15] etc. typically have three major modules such as image acquisition, feature extraction and matcher. System takes matching decision based on a predefined threshold which is chosen in such a way that the performance of the system is optimized. One such parameter is the equal error rate (EER) where false accept rate (FAR) and false reject rate (FRR) are same. Let us consider two independent biometric systems, S_1 and S_2 , with their score densities or matching score distributions,

$s_I(x)$ and $s_2(x)$, respectively. Score density represents the likelihood of the score for the matcher. Therefore, $\int_0^1 s_1(x) dx = \int_0^1 s_2(x) dx = 1$. A matching that involves biometric samples from the same user is called *genuine*; otherwise it is *imposter*. Let H_I and H_G be the hypothesis that the match is of type *imposter* or *genuine* respectively. So if $s(x|H_G)$ and $s(x|H_I)$ are the score densities for genuine and imposer matching then $\int_0^1 \{s(x|H_G) + s(x|H_I)\} dx = 1$. A typical integrated system situation is shown in Fig. 1, it involves three independent biometric systems A , B and C . It is shown that input and output to any module of independent system is not restricted to flow in the same system.

Let the threshold at EER for $S1$ and $S2$ be t_1 and t_2 respectively with $t_1 < t_2$. It can be shown that the threshold t for an integrated system involving $S1$ and $S2$ cannot be less than t_1 . Fig. 2 illustrates that false rejection rate ($FRR = \int_0^t s_1(x|H_G) dx$) decreases with decrease in threshold. For $S1$ and $S2$ we have $\int_0^t s_1(x|H_G) dx < \int_0^{t_1} s_1(x|H_G) dx$ and $\int_0^t s_2(x|H_G) dx < \int_0^{t_2} s_2(x|H_G) dx$. Also, since false acceptance rate ($FAR = \int_t^1 s(x|H_I) dx$) increases with decreases in threshold so $\int_t^1 s_1(x|H_I) dx > \int_{t_1}^1 s_1(x|H_I) dx$ and $\int_t^1 s_2(x|H_I) dx > \int_{t_2}^1 s_2(x|H_I) dx$. As according to assumption t_1 and t_2 are thresholds at the point of equal error rate, so false acceptance and false rejection rates are also equal there (i.e. $FAR = FRR$). Alternatively, it can be said that for $S1$ and $S2$, we have $\int_0^{t_1} s_1(x|H_G) dx = \int_{t_1}^1 s_1(x|H_I) dx$ and $\int_0^{t_2} s_2(x|H_G) dx = \int_{t_2}^1 s_2(x|H_I) dx$. This implies, $\int_0^{t_1} s_1(x|H_G) dx + \int_0^{t_2} s_2(x|H_G) dx = \int_{t_1}^1 s_1(x|H_I) dx + \int_{t_2}^1 s_2(x|H_I) dx$. Using the inequalities, we can conclude that $\int_0^t s_1(x|H_G) dx + \int_0^t s_2(x|H_G) dx < \int_t^1 s_1(x|H_I) dx + \int_t^1 s_2(x|H_I) dx$. Since this holds for any $t < t_1$, we cannot get EER for integrated system below the threshold t_1 .

Similar arguments can also show that the threshold for the integrated distribution cannot be greater than t_2 . At a threshold between t_1 and t_2 , the errors would be sum of errors for each score distribution. For $S1$, false accepted rate FAR_{S1} is decreased by $\delta_{FAR_{S1}}$ while false reject rate FRR_{S1} is increased by $\delta_{FRR_{S1}}$. Similarly, for $S2$, false accepted rate FAR_{S2} is decreased by $\delta_{FAR_{S2}}$ with the increase of false reject rate FRR_{S2} by $\delta_{FRR_{S2}}$. Therefore, for a integrated system, $FAR_{integrated} = FAR_{S1} + \delta_{FAR_{S2}} + FAR_{S2} - \delta_{FAR_{S2}}$ and $FRR_{integrated} = FRR_{S1} - \delta_{FRR_{S1}} + FRR_{S2} + \delta_{FRR_{S2}}$.

3 Proposed Technique

This section describes the proposed normalization based technique for improving the accuracy of integrated system when multiple independent matchers are involved. It can be shown that any uniform linear operation like adding or multiplying some fixed.

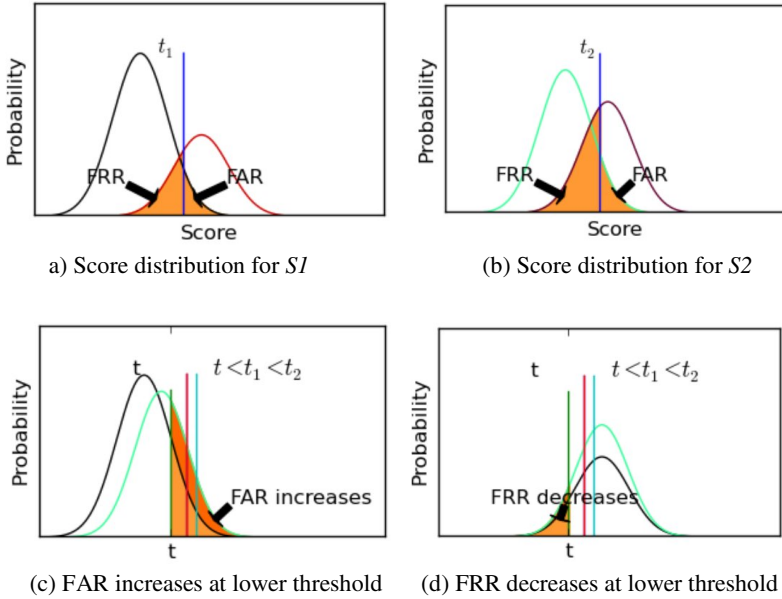


Fig. 2. Effect of threshold lowering on FAR and FRR

value to all the matching scores does not effect the performance of the system. Consider a system S having score distribution as s , false accept rate (FAR) and false reject rate (FRR) are defined as $FAR = \int_t^1 s(x|H_I) dx$ and $FRR = \int_0^t s(x|H_G) dx$. By adding δ to all the matching scores, new distribution s' becomes $s'(t)=s(t-\delta)$. Therefore, new false accept rate and false reject rate are $FAR' = \int_{t+\delta}^1 s'(x|H_I) dx = \int_{t+\delta}^1 s(x-\delta|H_I) dx$ and $FRR' = \int_0^{t+\delta} s'(x|H_G) dx = \int_0^{t+\delta} s(x-\delta|H_G) dx$. Since $\int_{t+\delta}^1 s(x-\delta|H_I) dx = \int_t^1 s(x|H_I) dx$. Therefore, $FAR' = FAR$ and $FRR' = FRR$. This puts an upper bound on the value of FAR and FRR of the integrated system. Consider an integrated system composed of two independent systems, $S1$ and $S2$, then FAR and FRR of the integrated system are given by $FAR_{Integrated} = \frac{FAR_{S1}+FAR_{S2}}{2}$ and $FRR_{Integrated} = \frac{FRR_{S1}+FRR_{S2}}{2}$ respectively.

Different independent matchers produce their output scores in different range. Therefore, an initial strategy is required to unify the score range. One such scheme is to divide the output score by the maximum possible range of the matcher. In other words, for a matcher score in the range between 0 and Max_Score is obtained by dividing by Max_Score.

$$Matching\ Score = Score\ Output / Max_Score \tag{1}$$

Decision threshold of the integrated system is different from the threshold of individual system. To specify appropriate new threshold in the integrated environment, we adopt a registration-correction approach which basically modifies the matching



Fig. 3. Scan of same finger on different fingerprint scanners

scores such that it pegs the threshold of individual system to the threshold of the integrated system. All participating individual systems are needed to be registered a priori.

A target threshold T_0 for the integrated system is initially chosen in the range $[0,1]$ to peg the other systems. This value is used as the threshold for every matcher of the registered systems for genuine/imposter decision while matching. New independent biometric system to be included in the integrated system is first registered to determine its correction parameters. Registration of a new system S involves the creation of a small dataset D_S of identified subjects using the scanner of S . Let the threshold at which equal error rate EER_S of the system occurs be T_S . Correction factor for the matcher S is the value $N_S = T_0 - T_S$. All output matching scores of the system S for two feature templates, are applied to Equation (1) and then correction factor N_S is added to normalize the score. Decision threshold used for the system S to conclude for genuine/imposter decision is set to be T_0 . It has been shown that neither the division using Equation (1) nor the addition of the fixed value N_S to all output scores have any impact in the matching performance of the individual system S . But it improves the overall performance of the integrated system.

4 Experimental Results

This section discusses the experimental setup, database used and the experimental results in detail. Subsection 4.1 describes specification of the devices used in the setup. The database used in this paper is discussed in Subsection 4.2 and experimental results are analyzed in Subsection 4.3.

4.1 Experimental Setup

The technique has been analyzed by using three well known biometric systems. To maintain the anonymity let us refer them as A, B and C. Each system comes with a

Table 1. Biometric System Specifications

Specification	System A	System B	System C
Image Resolution	500 DPI	500 DPI	500 DPI
Image Size	352x544 pixels	260x300 pixels	320x480 pixels
Image Gray Scale	8-bit gray level	8-bit gray level	8-bit gray level
Effective Sensing Area	17.5x27.5 mm ²	12.7x14.9 mm ²	16.26x24.38 mm ²
Light Source	Blue LED	Red LED	Infrared LED
Template size	Upto 5000 bytes	Upto 400 bytes for ANSI and 800 bytes for ISO/IEC 19794-2	Upto 5000 bytes

fingerprint scanner and a software development kit (SDK) having modules for image acquisition, feature extraction and matching. Acquired images are stored in bitmap file format [16]. Technical specifications of the three systems are given in Table 1.

4.2 Database

The proposed technique has been analyzed using IITK-Rural database of 500 subjects. It consists of fingerprint images acquired with the help of three types of fingerprint scanners, viz. A, B and C. Data has been acquired from the rural areas under non-controlled environment in two sessions with a gap of two months between the two sessions. Each individual has given data of his/her finger three times on each scanner. Fig. 3 shows fingerprint images acquired on three different scanners.

The database consists of 3000 images. Randomly selected 100 subjects from IITK-Rural database are reserved for registration. The 600 images form a dataset D_T . Matching strategy is across the phases; therefore, the fingerprint acquired in first phase are matched with all the images acquired in second phase. So in total there are 90000 matchings for each matcher. Features from each image are extracted and stored in ISO format template. ISO/IEC 19794-1:2011 standards [17] specifies a biometric interchange format to transfer feature sets from one vendor to another by describing the general aspects and requirements. Its goal is to allow the interoperability among biometric recognition systems. This defines the representation of minutiae and storage from fingerprint ridge and valley patterns. It also mentions the representation format for the features.

4.3 Results

The matchers of independent system A, B and C use their proprietary algorithms to compute the similarity score and they have different ranges of the matching scores. Table 2 shows the observed range of matching scores for each matcher.

Table 2. Range of scores for matchers A, B and C

Matcher	Minimum	Maximum
A	0	100000
B	0	199
C	0	499

The equal error rates (EER) for the system A, B and C are calculated using their respective matchers and the three thresholds T_A , T_B and T_C corresponding to their EER value are determined using D_T as described in Subsection 4.2. Correction factors for the three system A, B and C are determined by using the formula $N_A = T_0 - T_A$, $N_B = T_0 - T_B$ and $N_C = T_0 - T_C$ for each matcher. The threshold at EER and the correction factors for the three systems are also shown in Table 3.

Table 3. Threshold and correction factors during registration

	Matcher A	Matcher B	Matcher C
Threshold	0.155	0.070	0.045
Correction factor	-0.065	0.020	0.045

Remaining images of 400 subjects are used for testing. Across phase testing strategy performs matching one image of first phase with all images acquired in second phase. So, 1200 images of first phase are matched with another 1200 images of second phase. Therefore, 1440000 matchings are considered in total. Matching scores for this integrated system without applying the correction factors is analyzed

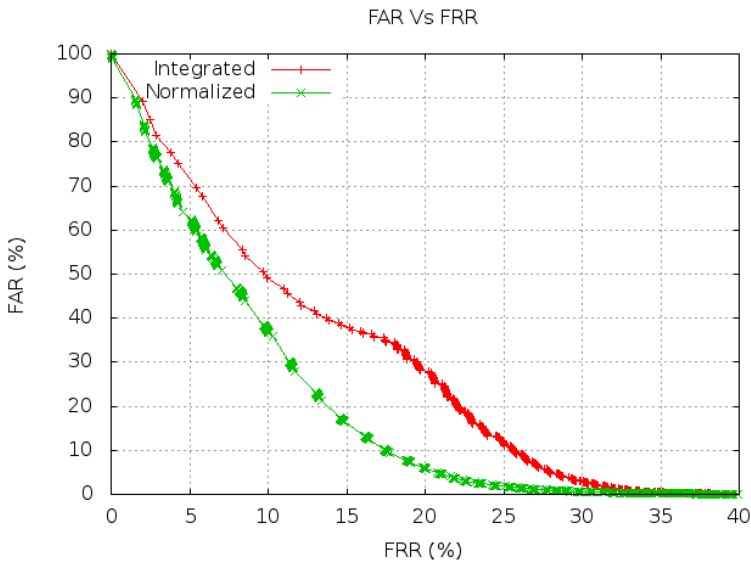


Fig. 4. ROC for integrated and normalized distribution

and it is observed experimentally that the value of equal error rate (EER) is 21.78% and the threshold is found as 0.110. Subsequently respective correction factor is applied on the output scores to get normalized one. Equal error rate (EER) values and threshold of normalized distribution are also determined. ROC for the naive integrated distribution from the three matchers and the proposed heuristic normalized distribution are shown in Fig. 4.

It can be observed from ROC curve drawn in Fig. 4 that for threshold at the extremes, ROC for the integrated distribution and the normalized distribution tend to behave similarly which is expected as the false acceptance or false rejection becomes large when threshold is closed to the extreme matching score. These are the regions where either FAR or FRR is closed to 100%. There is large difference between FARs of the two distributions when the FRR is lying between 10% to 25%. The maximum difference in FAR is found where the FRR value is 18%. It shows that the ROC curve for the normalized scores is lowered over a significant range of FAR and FRR; which implies that the normalized distribution lowers the error of the integrated distribution. EER for the normalized distribution is 15.66% and the threshold is found to be at 0.090. Thus the improvement in EER after normalization is found to be 6.12%.

5 Conclusions

The paper has proposed an heuristic based technique to improve the overall performance of an fingerprint based integrated biometric system. Advantage of the proposed system is that it does not requires modifying underline recognition approach of any independent biometric system. The technique registers every participating independent system at initial stage to assign correction factors and used later for pegging. Multi scanner IITK-Rural fingerprint database of 500 users is used which is collected in two sessions taking three fingerprint sample on three scanner per session. Experimental results reveal that it has improved the equal error rate (EER) of the integrated system from 21.78% to 15.66% by applying the heuristic based technique.

Acknowledgements. Authors like to acknowledge the support provided by the Department of Information Technology, Government of India to carry out this research work.

References

1. Chen, Y., Jain, A.K.: Beyond Minutiae: A Fingerprint Individuality Model with Pattern, Ridge and Pore Features. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 523–533. Springer, Heidelberg (2009)
2. Jain, A.K., Feng, J., Nandakumar, K.: Fingerprint Matching. *IEEE Computer* 43(2), 36–44 (2010)
3. Hong, L., Jain, A., Pankanti, S., Bolle, R.: Identity Authentication Using Fingerprints. In: Bigün, J., Borgefors, G., Chollet, G. (eds.) AVBPA 1997. LNCS, vol. 1206, pp. 103–110. Springer, Heidelberg (1997)

4. Ross, A., Jain, A.: Biometric Sensor Interoperability: A Case Study in Fingerprints. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 134–145. Springer, Heidelberg (2004)
5. Nigam, A., Gupta, P.: A New Distance Measure for Face Recognition System. In: International Conference on Image and Graphics, ICIG, pp. 696–701 (2009)
6. Nigam, A., Gupta, P.: Comparing Human Faces Using Edge Weighted Dissimilarity Measure. In: International Conference on Control, Automation, Robotics and Vision, ICARCV, pp. 1831–1836 (2010)
7. Kumar, J., Nigam, A., Prakash, S., Gupta, P.: An Efficient Pose Invariant Face Recognition System. In: Deep, K., Nagar, A., Pant, M., Bansal, J.C. (eds.) Proceedings of the International Conf. on SocProS 2011. AISC, vol. 131, pp. 145–152. Springer, Heidelberg (2012)
8. Bendale, A., Nigam, A., Prakash, S., Gupta, P.: Iris Segmentation Using Improved Hough Transform. In: International Conference on Intelligent Computing, ICIC, pp. 408–415 (2012)
9. Nigam, A., Gupta, P.: Iris Recognition Using Consistent Corner Optical Flow. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 358–369. Springer, Heidelberg (2013)
10. Singh, N., Tiwari, K., Nigam, A., Gupta, P.: Fusion of 4-Slap Fingerprint Images with Their Qualities for Human Recognition. In: World Congress on Information and Communication Technologies, pp. 925–930 (2012)
11. Singh, N., Nigam, A., Gupta, P., Gupta, P.: Four Slap Fingerprint Segmentation. In: Huang, D.-S., Ma, J., Jo, K.-H., Gromiha, M.M. (eds.) ICIC 2012. LNCS, vol. 7390, pp. 664–671. Springer, Heidelberg (2012)
12. Badrinath, G.S., Tiwari, K., Gupta, P.: An Efficient Palmprint Based Recognition System Using 1D-DCT Features. In: Huang, D.-S., Jiang, C., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2012. LNCS, vol. 7389, pp. 594–601. Springer, Heidelberg (2012)
13. Tiwari, K., Arya, D.K., Gupta, P.: Designing Palmprint Based Recognition System Using Local Structure Tensor and Force Field Transformation for Human Identification. *Neurocomputing* 6839, 602–607 (2012)
14. Nigam, A., Gupta, P.: Finger Knuckleprint Based Recognition System Using Feature Tracking. In: Sun, Z., Lai, J., Chen, X., Tan, T. (eds.) CCBR 2011. LNCS, vol. 7098, pp. 125–132. Springer, Heidelberg (2011)
15. Badrinath, G.S., Nigam, A., Gupta, P.: An Efficient Finger-Knuckle-Print Based Recognition System Fusing Sift And Surf Matching Scores. In: Qing, S., Susilo, W., Wang, G., Liu, D. (eds.) ICICS 2011. LNCS, vol. 7043, pp. 374–387. Springer, Heidelberg (2011)
16. MSDN: Bitmap Header Types, <http://msdn.microsoft.com/en-us/library/dd183386%28VS.85%29.aspx>
17. ISO/IEC 19794-2:2011 Information Technology – Biometric Data Interchange Formats – Part 2: Finger Minutiae Data (2011)

Edge Based Steganography on Colored Images

Mangat Rai Modi, Saiful Islam, and Phalguni Gupta

Department of Computer Science and Engineering,
Indian Institute of Technology, Kanpur
Uttar Pradesh - 208016, India
{mangat, sislam, pg}@iitk.ac.in

Abstract. This paper proposes a novel technique to hide secret messages in the least two significant bits of the edges in the cover image. Edges make a better option to hide secret data than any other region of an image where a small distortion is much more noticeable. Edge detection is adaptive to the amount of data in the message. Smaller is the amount of data, more striking edges are selected. Experimental results have shown that the technique performs better or at par with the existing steganography techniques against visual attacks.

Keywords: Steganography, Edge Detection, LSB Matching, Steganalysis, Information Hiding.

1 Introduction

Essence of image steganography is to hide data in a way that the third party should not be able to detect the presence of the secret message in the cover media. Note that, if the third party becomes aware of the presence of some secret message, it indicates that the steganography technique has failed to serve its purpose. Steganalysis is the technique of detecting the presence of secret message in the media under transmission. Whenever a message is hidden in the cover media (here cover image), it distorts some of the properties. This distortion may be visible to human vision system in stego image (either raw or processed). Steganalysis techniques exploit these distortions to detect the presence of secret messages. To prevent detection, a steganography technique should produce minimum distortion possible while hiding data. Edges are the areas in the image, where there is a sharp change in the visual property of the image. An edge pixel is different from the neighbouring pixels and this sudden change offers more resistance to distortion. Hiding data in them do not produce any visual anomalies. In this paper a novel steganography technique is proposed which hides message in the pixels belonging to the edges in an image. The technique decides the threshold filter depending on the length of data, so that more prominent edges are chosen. The experimental results have shown that the proposed technique works better or at par with the existing edge steganography techniques in terms of edge selection and security against visual attacks.

The paper is organised into following sections. Section 2 gives an overview of some relevant steganography and steganalysis techniques. Section 3 explains the proposed technique with the details of edge selection and embedding whereas all the experimental results are shown and explained in Section 4 and Section 5 concludes the paper with final remarks.

2 Literature Review of Previous Techniques

There exist several edge based steganography techniques. Among which, well known are the techniques based on Pixel Value Difference. In Pixel Value Difference (PVD)[1], an image is divided into non-overlapping blocks of adjacent pixels. These blocks are randomly selected by a Pseudo Random Number Generator (PRNG) and data is embedded into each of its pixels. Amount of data embedded, i.e. number of last significant bits used, is directly proportional to the difference in the values of adjacent pixels. This uneven embedding in PVD leads to unusual steps in the histogram of pixel difference in the stego image. Improved Pixel Value Difference (IPVD)[2] has exploited this vulnerability and another improved version was proposed. Adaptive Edge LSB (AE-LSB)[3] has also removed this uneven pixel difference by introducing a readjusting phase and has provided better capacity. All these techniques are edge adaptive in a way that they embed more data where pixel difference is high but they all have one fundamental limitation. They pick pixel-pair at random rather than picking on the basis of higher differences. So, they end up embedding data at random places in the image, distorting the texture in LSB plane of the image. Also these techniques show poor results when compared with other steganalysis techniques in [4]. There exist another edge based technique known as Hiding Behind Corners (HBC)[5], where a corner pixel is found by applying Laplace Formula. This pixel is then used to contain hidden data. Data is embedded by using simple LSB Substitution where bits in the LSB plane are flipped if they are not same as the message bits. Such embedding leads to many structural asymmetries (even valued pixels never decreases and odd ones never increase). These anomalies could be easily be detected by many structural steganalysis tools like Chi-square[6], RS steganalysis [7], Sample Pair analysis (SP)[8] and Weighted-stego (WS)[9] as shown by the authors in [4]. Thus HBC maintains texture in LSB plane but others poor security. Instead of LSB Substitution, LSB Matching (LSBM) is better for data embedding where instead of bit flipping, it adds +1 or -1 with equal probability to the pixels with LSB values different from the message bits. EALMR [4] is another edge based technique, which embeds data using modified LSBM based technique called LSB Matching Revisited (LSBMR). It calculates the difference between adjacent pixel values, and if it is greater than a predefined threshold, both the pixels are marked as part of an edge. Value of threshold is set high to get only the striking edges, which is decreased as it proceed further to accommodate whole message. This edge selection algorithm used in this technique is not sophisticated and can return false edges. Which means some prominent edges will not be detected and data will be embedded in smoother parts.

The steganography techniques discussed above cause very little distortion in the stego image which is invisible to naked human eye as such, but it could be noticeable if we look at the LSB plane of a stego image. LSB plane is the binary image formed by choosing Least Significant Bit of each corresponding pixel of the stego image. LSB planes for some cover images are not always random, but retain some texture. Texture is formed by the clusters of pixels with same LSB value, where we can see big white and black patches. Now, edge based steganography techniques refrain from any embedding in the smoother parts, hence maintain texture in LSB Plain. In this paper a novel steganography technique has been proposed, which works exclusively on colored images. It selects sharp edges based on length of secret message to accommodate data.

3 Proposed Approach

The proposed approach works on colored images. Edges are found using Canny Edge Detection algorithm [10] in one of the plane. Then the pixels in other two planes corresponding to the edge pixels are selected. These pixels will hold the bits of secret message. From the set of these chosen pixels, a random walk is generated using the password needed for decryption as the seed. This password, also known as stego-key is shared between sender and receiver of the secret message.

Input: I:Image, S: Size of Edge set required

Output: threshold: t_h for canny to get S pixels

```

1. // no. of edge pixels,  $n_e \leq S + 0.01 \times S$ 
2.
3. limit  $\leftarrow 0.01 \times S$ ;
4.  $t_{\max} \leftarrow 1$ ;
5.  $t_{\min} \leftarrow 0$ ;
6. threshold  $\leftarrow (t_{\max} + t_{\min}) / 2$ ;
7. // no. of edge pixels returned by canny ne
8.  $n_e \leftarrow \text{count}(\text{canny}(I, \text{threshold}))$ ;
9. difference  $\leftarrow n_e - S$ ;
10. while (difference > limit) or (difference < 0) do
11.
12. // we have found fewer or too many edges than
    required
13. if difference > limit then
14.      $t_{\min} \leftarrow \text{threshold}$ ;
15.     threshold  $\leftarrow (t_{\max} + t_{\min}) / 2$ ;
16. end
17. else if difference < 0 then
18.      $t_{\max} \leftarrow \text{threshold}$ ;

```

```

19.     threshold  $\leftarrow$  (  $t_{\max}$  +  $t_{\min}$  ) / 2;
20.     end
21.      $n_e \leftarrow$  count( canny( I, threshold));
22.     difference  $\leftarrow$   $n_e - S$ ;
23.     end

```

Algorithm 1. getThreshold: Get z_s value

The technique is adaptive to the amount of data to be hidden. Canny takes two types of thresholds to find the edges. High threshold (t_h) is used to identify the main edges and lower threshold (t_l) is used to further extend those edges found using t_h . In this technique t_l is taken as $4/10^{\text{th}}$ part of t_h . Higher is the value of t_h , more sharper edges are found using. It means, if we lower the value of t_h , more number of edges (less sharper though) will be found. So t_h value is adjusted to the point that just enough edges in the cover image are found to hold the secret message. Algorithm 1 (getThreshold) determines the value of t_h to get enough edges. So, the sharpest possible edges are obtained using canny to hold the message completely. Sharper edges provide sharp change in the visual properties in the image, so there is less chance of detecting hidden data. As edges provide more robustness, two bits of message are embedded in the two least significant bits of each edge pixel. Since the threshold is dynamically generated based on message length during embedding, it cannot be the part of shared stego key. It is hidden in a predefined area in the image, using IEEE-754 Half Precision Floating Point format to denote the value of threshold in binary. Now along with the length of data (in binary), 16 more bits (for threshold) are required to embed with the message. To hide a message of say N bits we require following fields:-

- Secret message = N bits
- Message length = $\log(N)$ bits
- Threshold value = 16 bits
- Total no. of pixels = $(N + \log(N) + 16) \times 4$

For data embedding we have used a modified LSB Matching approach. If the LSB of the edge pixel is not same as message bit, then similar to LSBM, we randomly add +1 and -1 to it. But in the case of second least significant bit, we add +2 or -2, keeping the final value in the range of [0-255] (we have assumed image to be of 24 bit color depth, i.e. each pixel in a plane is of 8 bit). As a result difference between original and modified pixel lies in the range of [-3, 3]. This whole embedding part is shown in Algorithm 2 (edgeSteg).

Input: I: Image, M: Secret Message, P: Password

Output: S: Stego image

1. $S \leftarrow I$;
2. // find edges in one plain and embed in other two

```

3. // r, g and b are red, green, blue plane
   respectively
4. r ← S(1);
5. g ← S(2);
6. b ← S(3);
7. M1 ← M(1..length(M)/2);
8. M2 ← M(length(M)/2+1..end);
9. threshold ← getThreshold(g,length(M)/4);
10. e ← canny(g,threshold);
11. e ← randomPermute(e,P);
12.
13. // e is set of indices of edge pixels randomly
   permuted by using P as seed
14. index ← 0;
15. for i ∈ e do
16.   if ri[lsb] ≠ M1index then
17.     //ri[lsb]: LSB of ith pixel of red plane
18.     // M1index: index bit of M1index
19.     // ±1 signifies that 1 could be added or sub-
       -tracted by equal probability
20.     ri[lsb] ← ri[lsb]± 1;
21.   end
22.   if ri[lsb - 1] ≠ M1index+1 then
23.     ri[lsb - 1] ← ri[lsb - 1]± 2;
24.   end
25.   if bi[lsb] ≠ M2index then
26.     bi[lsb] ← bi[lsb]± 1;
27.   end
28.   if bi[lsb - 1] ≠ M2index+1 then
29.     bi[lsb - 1] ← bi[lsb - 1]± 2;
30.   end
31.   index ← index + 2;
32. end
33. end
34. S(1) ← r;
35. S(2) ← g;
36. S(3) ← b;

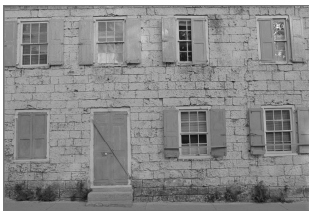
```

Algorithm 2. edgeSteg: Hides message

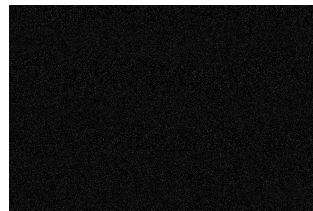
4 Experimental Results

The proposed technique has been tested over the database of 2142 images. These images are part of NRCS database [11] which are originally scanned from print media to be stored as 32 bit CMYK TIFF format in digital form. These are converted into 24

bit RGB profile by using color profiles available at [12] and are cropped to 2100x1500 resolution. Secret message is randomly generated by a pseudo random generator (PRNG) at run time. Message length is such that only 10% of the cover image pixels has been used to contain message. After embedding the proposed technique is then compared with LSBM, HBC and EALMR for security against visual attacks. The resulted stego images from all the techniques were saved in loseless BMP format. The proposed technique, HBC and EALMR embed data in the edges to prevent loss of texture in LSB plane. To get more insight on edge selection, difference between red plane of stego images and that of the cover images has been calculated. Fig. 1 is an example output where modified pixels are denoted by white color. We can see that LSBM hides data without any discrimination to edges or smoother parts. However HBC and EALMR clearly modify pixels along the edges. However in both techniques, some prominent edges are left while data is instead embedded in smoother parts. For example, edges around bricks were left, and data was filled inside bricks in the given image. In Fig. 1(e) one may see well defined white edges and smoother black regions.



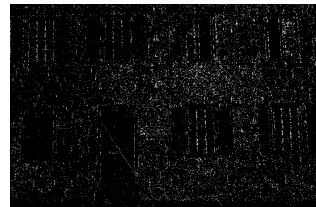
(a) Red Plane of the cover image



(b) LSBM



(c) HBC



(d) EALMR



(e) The Proposed Technique

Fig. 1. Difference of red plane of cover image with that of stego images each having 10% of data embedded

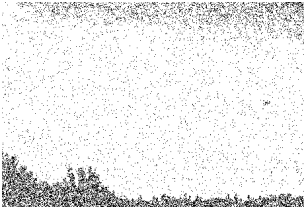
One of the reasons to use edge based steganography is to preserve the texture in LSB plane. To understand the meaning of texture in LSB plane consider Fig. 2(b) where one can see some white and black patches. These are the clusters of pixels with same LSB value. This property can be generally seen in smoother areas of image like bright sky or dark shadow. So a change in smoother parts of the image may change LSB value of the pixels in these clusters. This effect is seen in Fig. 2(c) which is the zoomed part of the stego image generated by LSBM, corresponding to the area marked in Fig. 2(b) where LSBM writes some message bits causing some black pixels to appear. These black pixels on the white patch may raise some suspicion. On the other hand, all the Edge based techniques in Fig. 2(d),(e) and (f) do not cause any noticeable change.



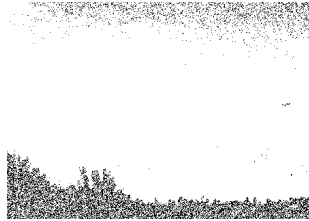
(a) Cover Image



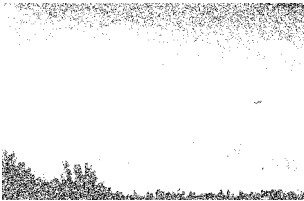
(b) LSB Plane of cover Image



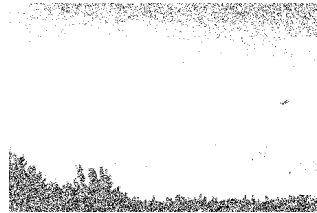
(c) LSBM



(d) EALMR



(e) HBC



(f) Proposed Technique

Fig. 2. LSB plane of red plane using (a) as cover and 10% embedding. Marked smooth area is cropped from the stego images.

5 Conclusions

This paper has proposed a new data hiding technique in color images. Data is hidden at the edges of cover image selected using a threshold adaptive to the message length. Against visual attacks the proposed technique, along with EALMR and HBC provides more robustness than LSBM where black and white dots were created in the smoother parts of stego image. However the proposed technique chooses more striking and well defined edges because of its better edge selection algorithm as compared to both EALMR and HBC.

Acknowledgements. Authors like to acknowledge the support provided by the Department of Information Technology, Government of India to carry out this research work.

References

1. Wu, D., Tsai, W.: A Steganographic Method for Images by Pixel Value Differencing. *Pattern Recognition Letters* 24(9-10), 1613–1626 (2003)
2. Zhang, X., Wang: Vulnerability of Pixel-Value Differencing Steganography to Histogram Analysis and Modification for Enhanced Security. *Pattern Recognition Letters* 25(3), 331–339 (2004)
3. Yang, C., Weng, C., Wang, S., Sun, H.: Adaptive Data Hiding in Edge Areas of Images with Spatial Lsb Domain Systems. *IEEE Transactions on Information Forensics and Security* 3(3), 488–497 (2008)
4. Luo, W., Huang, F., Huang, J.: Edge Adaptive Image Steganography Based on Lsb Matching Revisited. *IEEE Transactions on Information Forensics and Security* 5(2), 201–214 (2010)
5. Hempstalk, K.: Hiding Behind Corners: Using Edges in Images for Better Steganography. In: *Proceedings Computing Womens Congress*, Hamilton, New Zealand (2006)
6. Westfeld, A., Pfitzmann, A.: Attacks on Steganographic Systems. In: Pfitzmann, A. (ed.) *IH 1999. LNCS*, vol. 1768, pp. 61–76. Springer, Heidelberg (2000)
7. Fridrich, J., Goljan, M., Du, R.: Detecting Lsb Steganography in Color and Gray-Scale Images. *IEEE MultiMedia* 8(4), 22–28 (2001)
8. Dumitrescu, S., Wu, X., Memon, N.: On Steganalysis of Random Lsb Embedding in Continuous-Tone Images. In: *International Conference on Image Processing*, vol. 3, pp. 641–644 (2002)
9. Ker, A.D., Böhme, R.: Revisiting Weighted Stego-Image Steganalysis. In: *Proceedings SPIE 6819, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, San Jose, CA, pp. 681905-1–681905-17 (2008)
10. Canny, J.: A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
11. Nrcs photo gallery (2013), <http://photogallery.nrcs.usda.gov/res/sites/photogallery/>
12. Adobe color profiles (2013), <http://www.adobe.com/support/downloads/detail.jsp?ftpID=3680>

An Evolutionary Optimization Method for Parameter Search in 3D Points Cloud Reconstruction

Vitoantonio Bevilacqua^{1,3,*}, Fabio Ivona^{1,2},
Domenico Cafarchia¹, and Francescomaria Marino^{1,4}

¹ Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari,
via Orabona 4 – 70126 Bari – Italy

² Trait d'Union srl,

via R. Redi 3 – 70124 Bari – Italy

³ eBIS srl,

via G. Petroni 25 (15/F4)– 70125 Bari – Italy

⁴ APIS Apulia Intelligent Systems srl,

via P. Fiore 26 – 70125 Bari – Italy

bevilacqua@poliba.it

Abstract. Reconstruction of 3D laser scanned point clouds may generate a mesh characterized by a high number of triangles. Unfortunately, in Computer Aided Design environments neither a simple triangle reduction, nor decimation filters are feasible for mesh optimization, because of their intrinsic errors.

In this paper we show how Genocop III can be effectively used to reconstruct a point cloud bounding the error under a certain threshold. Moreover, we define an optimized algorithm for evaluating the reconstruction error, that exploits AABB-trees and pre-computation and provides a useful metric to the genetic algorithm.

Keywords: 3D Point Cloud, Reconstruction, Decimation, Hole Filling, Genetic Algorithms, Evolutionary Computation, Reconstruction Error, AABB-trees.

1 Introduction

Laser scanning of 3D surfaces allows capturing huge points cloud datasets that can be used in a Computer Aided Design (CAD) environments. After a preliminary data-cleaning and registration phase, a digital representation of the original surface can be computed through a process of surface reconstruction that generates a polygonal mesh, usually made of triangles. However, reconstructing a surface for further use in architectural CAD software introduces two more requirements:

- reconstructed mesh triangles count should be as lowest as possible;
- maximum error caused by reconstruction should be bounded by a pre-defined threshold.

* Corresponding author.

Usually, a surface optimization algorithm aims at reducing the triangles count to a specific value, while binding the consequent error under a specified threshold. In order to make a surface reconstruction algorithm feasible for architectural CAD environments, instead, we are mainly interested in minimizing the triangles count, while maintaining the maximum reconstruction error under a specified threshold.

Such a result is usually obtained by carefully searching and setting reconstruction and post-processing parameters. In this paper we investigate how to exploit genetic algorithms in order to automatically set the parameters, guaranteeing a low complexity mesh, paying a feasible error. Despite of the high efficiency of modern reconstruction techniques, an external hint is still required in order to point out the parameters used to generate and simplify the output mesh: our main contribution is to define an efficient and automatic algorithm to find the best ratio between reconstruction error and triangle count. Additionally, our algorithm allows the integration of different reconstruction, decimation and hole filling methods, helping to find the best configuration for the reconstruction system.

Moreover, since to define a way to compute the reconstruction error is not a trivial task (and it turns to be even more difficult, when executed within a genetic algorithm iteration), we propose a way to solve the 3D distance problem by means of a fast 2D algorithm, optimized for a huge number of iterations, which takes advantage of search trees and point-triangle distance techniques.

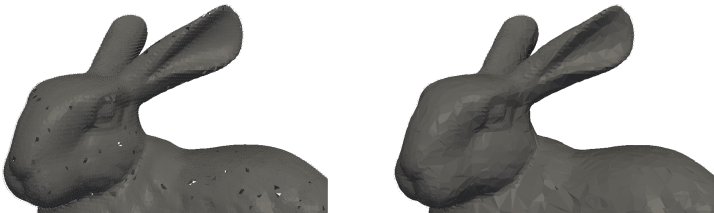


Fig. 1. Mesh reconstruction and optimization from 71787 to 24161 triangles

The paper is organized as follows. In section 2, we outline some mesh reconstruction, hole filling and decimation procedures that have included in our algorithm. Genocop III, an evolutionary genetic algorithm that handles non-linear constraints is then shortly recalled. In section 3, we describe the implementation of our algorithm for 3D surfaces reconstruction optimization. In section 4, we define a metric to provide an accurate estimation of the error generated by the reconstruction and decimation process. Experimental results, carried out both on syntetical benchmark and real world point clouds, are presented in section 5, evidencing the ability of the system to optimize the reconstructed mesh by finding the best parameter set. Conclusions are pointed out in section 6.

2 Background

Reconstruction and post-processing, starting from a points cloud, can be realized with a great number of algorithms, each characterized by strengths and weaknesses.

The choice of the algorithms and their parameters is usually committed to human operators, and this phase may require an intensive tuning step, in order to obtain an output mesh that minimizes the triangles count by satisfying the maximum error requirements.

2.1 3D Reconstruction

A typical surface reconstruction algorithm takes as input noisy points clouds that do not encode information on the original surface topology. We have considered the CPU-time and the quality of reconstruction as discriminating factors for the 3D reconstruction algorithm to adopt in our surface reconstruction procedure. This choice is due to the subsequent use of a genetic algorithm, which typically suffers time steps. In the following, we will explain the techniques that underlie the main steps of our surface reconstruction procedure. They are preliminary to the adoption of the genetic algorithm: in particular, after the reconstruction phase, the hole filling and the decimation filtering phases are necessary in order to achieve a low triangles count.

In [1], Hoppe proposes an approach based on a function that estimates the signed distance of each point from an unknown surface S . The key operation to define the distance function is to associate an oriented tangent plane with each point. The tangent plane T_i associated with the point x_i is built by a point o_i (center) and a normal vector n_i . Under this assumption, the distance from an arbitrary point p_k of the cloud to the plane T_i equals to:

$$dist_{T_i}(p_k) = (p_k - o_i) \cdot n_i$$

The subsequent phase aims to find geometrically close points, by checking if their corresponding tangent planes are consistently oriented. Starting from these ones, the distance function of p_i to an unknown surface can be computed using the oriented tangent planes: first, the algorithm finds a tangent plane T_i whose center o_i is the closest to p_i . The tangent plane is a local linear approximation of the surface S , so the signed distance $f(p)$ can be computed as the distance of p_i from the tangent plane T_i . The real surface can be described as the zero-set of the signed distance function. In a last stage, a marching cubes based contour tracing algorithm is used to approximate the zero set with a triangular mesh. The algorithm leads to the creation of an extremely dense initial mesh, that needs to be optimized in order to be feasible in a computer graphics environment. In order to achieve this result, Hoppe defines an automatic optimization procedure that aims to reduce the mesh complexity. The main target is to find a simplicial complex K and a set of vertex positions V that define a new mesh $M' = (K, V)$ by minimizing an energy function $E(K, V)$:

$$E(K, V) = E_{dist}(K, V) + E_{rep}(K) + E_{spring}(K, V)$$

where E_{dist} (distance energy) equals to the sum of the quadratic errors from the cloud to the mesh, and is calculated by identifying the nearest triangle to each point

and computing their distance; E_{rep} (representation energy) is introduced in order to penalize meshes with a high number of vertexes, and is proportional to a penalization factor defined by the user; E_{spring} (spring energy) is useful to ensure the convergence of the optimization algorithm.

In [2], Kazhdan et al. face the reconstruction problem by defining a 3D indicator function χ , defined as 1 for the points inside the surface and 0 for those outside. As the indicator function gradient is a vector field that has a non-zero value only in proximity of the surface, computing the indicator function reduces to the search for the scalar function χ whose gradient best approximates the vector field \mathbf{V} defined by the points cloud. Applying the divergence operator, this variational problem is transformed into a standard Poisson problem of *computing the scalar function χ , whose Laplacian equals to the divergence of the vector field*

$$\mathbf{V}: \delta_\chi \equiv \nabla \cdot \nabla_\chi = \nabla \cdot \mathbf{V}.$$

A different approach to the reconstruction problem is defined by the algorithm Greedy Surface Triangulation [3], based on the concept of surface growing. The mesh generation algorithm proceeds incrementally by searching, for each point P_i , a k -neighborhood made by the k nearest point from P_i within a sphere of radius $r = \mu \cdot d_0$, where d_0 is the distance from P_i to his closest point, and μ is a user specified constant. The neighborhood is then projected on a plane that is approximately tangential to the surface formed by the neighborhood. Projected points, whose visibility is occluded from P_i by the mesh edges are removed, and the remaining ones are triangulated, obtaining the final mesh. As this greedy approach proceeds incrementally, without deleting edges, the new surface grows directly leading to the final mesh, avoiding the memory consumption of the two other approaches.

2.2 Hole Filling and Decimation

The reconstruction process may cause the presence of holes in the mesh, so a hole filling filter is required in order to reduce the reconstruction error during the subsequent decimation phase. As in [4], the adopted hole filling algorithm proceeds by detecting boundary edge rings and associates neighborhood of points. Then, the neighborhood is projected on a tangent plane, and new vertexes are interpolated. In the last step, new triangles are computed with the Moving Least Squares algorithm [5].



Fig. 2. Mesh reconstruction and optimization from 93836 to 20795 triangles

Afterward, triangles decimation is performed. Decimation is based on the approach described in [6], that chooses which vertexes can be removed by classifying them in simple, complex and boundary. A simple vertex is surrounded by a complete triangle cycle, and can also be identified either as an interior vertex (if it is part of two edges that generate a large angle between the triangles they belong to, named feature edge) or a corner vertex (if it belongs to one or more than two feature edges). A boundary vertex is surrounded by a half cycle of triangles and a complex vertex is surrounded by a complete cycle of triangles and belongs to a triangle not in the cycle. Every vertex that is not a complex nor a corner vertex, can be removed and put into a priority queue, ordered by crescent values of the error implied by their removal. Finally, the queue is processed removing each vertex and re-triangulating the generated holes. This process is repeated until an optimization target is met.

2.3 Genetic Algorithms and Genocop III

Genetic Algorithms [7][8] are complex adaptive procedures, aimed to solve optimization problems in several real world applications, and are based on natural evolution principles. They work by selecting the best solutions for a given optimization problem, recombining them to build new generation, and converging towards the best solution. Genetic algorithms usually follow these main steps:

1. an initial set of possible solutions is defined;
2. each solution is evaluated, and the best ones (in the sense of a given fitness function) are selected;
3. a new set of solutions is defined by manipulating the best solutions of the previous set. By this way, a good solution has better chances to reproduce itself, and to continue the evolution process. New solutions are obtained through mutation and crossover operations;
4. if a maximum iteration count is met, or the algorithm reaches an optimum solution, the optimization process ends. If not, the iterative process continues from step 2.

Usually, each solution (phenotype) is codified as binary code in chromosomes made of set of bits (genes). The evaluation of each phenotype is made through a fitness function that describes its attitude to solve the problem. Genetic Algorithms play a main role in the process of defining the best set of reconstruction and decimation parameters. Additionally, they ensure that the parameter search converges towards the best solution faster than other optimization techniques. We have chosen to adopt Genocop III [9][10] (GENetic algorithm for NUMerical OPTimization of CONstrained Problems), which supports nonlinear constraints, and is based on the concepts of co-evolution and repair algorithms.

3 Optimized Reconstruction

Our target is to find a parameter set that allows to obtain a reconstructed mesh characterized by the lowest triangles count, and leads to an error no higher than the defined

error threshold. This can be formalized as an optimization problem with a single non-linear constraint, the maximum error.

In order to exploit the computational efficiency of Genocop III, we have defined two functions that can represent the evaluation function and the non-linear constraint. Chromosomes modeling is a propaedeutic step to the entire optimization process, and allows to apply mutation and crossover operators, in order to combine the best individuals and to obtain a population that converges towards the optimization target. The optimization process starts from an initial population random solutions. Each individual is uniquely identified by a specific value set for reconstruction and post-processing parameters, needed for the mesh generation. Each parameter has its own domain, defined in Genocop III configuration file. The evaluation function returns the triangles count of the reconstructed mesh. Therefore, this forces the algorithm to compute a reconstruction function M for each reference point. Given a reconstruction and decimation parameter set, tied to the current reference point \bar{R} , the evaluation function can be defined as:

$$eval(\bar{R}) = M(\bar{R}).triangle_count$$

In order to bind the reconstruction error below the threshold, we have defined a non-linear constraint that takes advantage of the error function described in section 4. For a search point \bar{S} , each point of the original cloud is evaluated, measured its distance from the reconstructed mesh, and obtained the maximum quadratic error:

$$e_{\max}^2(\bar{S}) = \max\{dist(x_i, M(\bar{D}))\}$$

e_{\max}^2 is then used by Genocop III to evaluate if a search point \bar{S} can be feasible and, thus, entered in the reference population.

3.1 Optimization Workflow

The optimization algorithm can search for the best reconstruction parameter set P_R , as well as the best decimation parameter set P_D . The optimization process is modeled as a unique execution of the genetic algorithm, in which each search point $S = \{P_R, P_D\}$ contains reconstruction and decimation parameters. At each evaluation, the algorithm launches the mesh reconstruction function and computes the decimation filter. The main target of this optimization problem is to find a parameter set (\bar{P}_R, \bar{P}_D) such that

$$M(\bar{P}_R, \bar{P}_D).triangle_count = \min_{(P_R, P_D)} \{M(P_R, P_D).triangle_count\}$$

$$\wedge$$

$$\max_i \{dist(x_i, M(\bar{P}_R, \bar{P}_D))\} < user_defined_error$$

4 Error Computation

While filtering reconstructed meshes, our optimization algorithm needs an estimation of the output surface quality. In order to make the generated mesh feasible for a computer aided design environment, a reconstruction algorithm must guarantee that the maximum error deriving by the decimation process does not exceed the specified threshold; under this assumption a simple error estimation, as the one used in common decimation algorithms, is not satisfactory. To face this problem, we have defined an error computation function that could be used by the genetic algorithm to check whether the generated population satisfies the maximum error constraint. This error metric is defined as the greatest distance between the cloud points and the generated geometry, i.e. the maximum deviation of the reconstructed mesh from the input point cloud. In order to compute the distance of a single point from the mesh, the error algorithm searches for the nearest triangle through an AABB-tree and a “point-triangle distance” is then computed.

As described by Jones in [11], the point-triangle distance could be obtained by projecting the point, say P , onto the plane of triangle T , and evaluating the position of the projection P_T . If P_T lies inside the triangle, the length $|PP_T|$ is the distance of P from T . If instead P_T falls outside the triangle, the point-triangle distance equals to the distance from P to the closest edge or vertex to P_T , depending of the projection position.

Starting from the optimized 2D method proposed by Jones, we implemented a point-triangle distance algorithm, that converts the problem into a two dimensional one and exploits pre-computation to achieve better computational performances. Our algorithm pre-computes a translation and a rotation matrix, to place the triangle so that V_1 lies on the origin, V_2 lies on the z axis and V_3 lies in the yz plane.

By this way we can reduce a 3D distance calculation into a bi-dimensional one, avoiding cross and dot products, and lowering the total number of computations needed to obtain the point-triangle distance. Transformation matrices are applied to translate and rotate the triangle and then are applied again to the point. P is then projected onto the plane of the triangle by setting its x coordinate to zero, obtaining P' .

Even though this process introduces an additional complexity to the distance computation process, transformation matrices allow to reduce the number of steps needed to obtain the distance, with computation times 94% lower than a matrix-free approach.

P' position is evaluated by computing its barycentric coordinates (u,v) respect of the translated triangle $V_1V_2V_3$. P' lies inside the triangle if

$$0 \leq u \leq 1, \quad 0 \leq v \leq 1, \quad u + v \leq 1$$

In this case, the point-triangle distance is $d = |P.x|$. If P lies outside the triangle, the point-triangle distance equals to the distance of P from the nearest of the triangle edges. For the edge V_1V_2 , its distance from P is computed by calculating the normalized projection of P' on it:

$$t = \frac{\overline{P'V_1} \cdot \overline{V_1V_2}}{\| \overline{V_1V_2} \|^2}$$

If $0 \leq t \leq 1$, the projection P'' lies within the segment $\overline{V_1V_2}$ and the point-triangle distance is $\sqrt{\| \overline{P'P''} \|^2 + |P.x|^2}$. If the projection P'' lies outside the triangle, the point-triangle distance equals instead to the distance from P to the nearest vertex (V_1 if $t < 0$ or V_2 if $t > 1$). Following the same process, the distance of P from V_1V_3 and V_2V_3 can be computed.

4.1 Error Computation Optimization

This phase requires two steps. A first step aims at reducing the point triangle distance CPU-time by applying a pre-computation process in order to reduce redundant calculations. A subsequent step allows a high decrease of the total number of distance evaluations by exploiting a binary search tree. By this way, for each point, the nearest triangle is identified and unnecessary distance computations are removed from the error estimation process.

Pre-computation: As more than one point usually share the same nearest triangle, the error algorithm can be optimized by pre-computing the transformation matrices as they depend on the triangle vertices, and remain constant when changing P coordinates. Additionally, barycentric coordinate equations

$$u = \frac{(V_2.y - V_3.y)(P'.z - V_3.z) + (V_3.z - V_2.z)(P'.y - V_3.y)}{(V_2.y - V_3.y)(V_1.z - V_3.z) + (V_3.z - V_2.z)(V_1.y - V_3.y)}$$

$$v = \frac{(V_3.y - V_1.y)(P'.z - V_3.z) + (V_1.z - V_3.z)(P'.y - V_3.y)}{(V_2.y - V_3.y)(V_1.z - V_3.z) + (V_3.z - V_2.z)(V_1.y - V_3.y)}$$

contain only four terms, depending on P' , that need to be calculated at each iteration. Other terms are constant and can be reduced to a constant factor, reducing the complexity of the computation.

Search trees: Another optimization factor is introduced by a nearest triangle search algorithm. During the pre-computation step, the mesh is organized upon a hierarchic tree (AABB-tree [12]) whose leaf nodes contain a single triangle each. The partitioning process works by iteratively subdividing the space, binding each triangle to the partition that contains the center of its Axis Aligned Bounding Box (AABB). The subdivision process continues until the whole search tree is defined. Because of the AABB-tree depth, a hint mechanism has been adopted in order to allow to begin the search from an intermediate node, instead of the root node. A kd-tree [13] is built together with the AABB-tree, and is used to identify the group of triangles that likely

contains the nearest one from the computed point. Because of the construction technique used, the triangles group identified by the kd-tree search is contained in the same AABB-tree node. By this way, it is possible to start the AABB-tree search from an intermediate node, reducing the computational complexity of the search algorithm.

Using search trees, our algorithm can evaluate the maximum error by computing a single distance for each point of the cloud, reducing the total number of point-triangle distance computations.

5 Test and Results

The optimization procedure has been verified using the reconstruction algorithm Greedy Projection Triangulation [3], because it is faster and more efficient than the other solutions analyzed in section 2. Even if we focused our study on a set of three reconstruction methods, our optimization algorithm ensures a high compatibility with every reconstruction and post-processing technique, and can adapt itself to the user requests. The algorithm has been then tested on benchmark point clouds, managing to achieve a good triangle count reduction, without exceeding the maximum error threshold. An initial execution of the algorithm on the Stanford Bunny dataset, shows the optimization rate in respect of the maximum error imposed (Fig. 3).

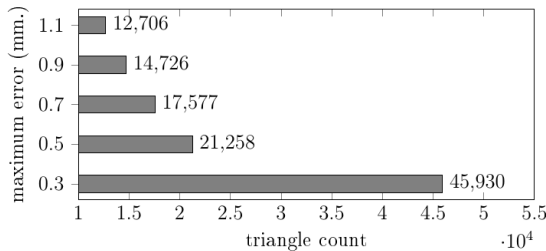


Fig. 3. Optimization rate

The maximum error of 0.3 mm is a lower threshold. Under this error value, our algorithm cannot optimize the mesh without breaking the precision requirement. Results for an error threshold of 0.5 mm are shown in Figure 1. Similar tests have been performed on different point cloud datasets, with increasing complexity. Our algorithm has been proved to be able to generate an optimized mesh with a low triangle count. Figure 2 evidences the presence of some reconstruction errors, due to the lack of points in some areas of the original cloud. The optimization algorithm interpolated the missing surface and applied the decimation filter in order to reduce by 80% the number of triangles. Figure 4 shows the reconstruction and decimation process starting from a high density point cloud. As it can be noticed, the reconstructed surface quality is high, with a maximum error lower than 0.2% of the point cloud width.

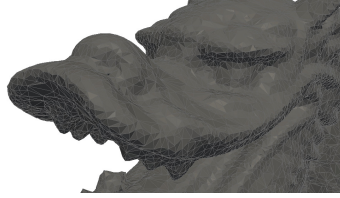


Fig. 4. Mesh optimization from 437645 points to 98857 triangles

A final test has been executed on a real world Trait d'Union company case project consisting of a 65506 points cloud scanned from the $3 \times 6 \times 5$ m³ environment of the *Auditorium Unità d'Italia, Isernia (Italy)*. The resulting dataset is characterized by not equally distributed points, huge areas without any information and a strong noise caused by the presence of a building yard. Starting from the point cloud, a very dense mesh has been reconstructed, obtaining 518626 triangles. The decimation algorithm managed to reduce the triangles count up to 186132, with a maximum error of 5 cm (Fig. 5).

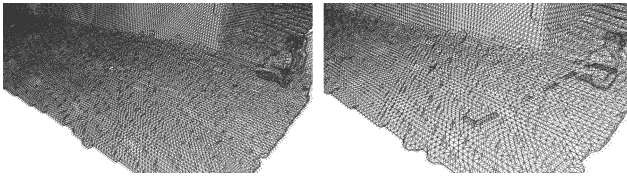


Fig. 5. Mesh reconstruction and optimization from 518626 to 186132 triangles

6 Conclusions

We have shown how genetic algorithms can be exploited in order to automatically obtain optimized mesh reconstruction parameters. The reconstruction process can fit within a computer aided design environment, defining an algorithm that complies with precision and triangle count constraints. Mesh optimization has proven effective with both low and high resolution point clouds, even if the resulting mesh quality is tied to the number of points in the original cloud and the maximum distance between them. In practical use, surface sampling with laser scanner introduces an intrinsic error that cannot be avoided, and it has to be taken into account, when defining the algorithm target error threshold. The main problem we had to deal with, was the high computational complexity of reconstruction algorithms that, in a genetic algorithm, become a strong factor of time consumption and get even worse when the number of points in the cloud increases. Starting from these assumptions, it is possible to define some areas of future research and development:

- further optimization the error computation algorithm with a better exploitation of search trees;
- point cloud partitioning, in order to execute the algorithm on smaller datasets and merge the optimized meshes;

- implementing a reconstruction and decimation algorithm oriented for an execution in a genetic algorithm;
- 3D geometric feature extraction and modeling of surfaces by using in a combined way genetic algorithms and Hough transform [14];
- porting the algorithm in a parallel environment (e.g., GP-GPU, FPGA).

Acknowledgements. This study was supported by the Italian PON FIT Project called “Sviluppo di un sistema di rilevazione della risonanza (SS-RR) N° B01/0660/02/X17” - Politecnico di Bari and AMT Services s.r.l. – Bari – Italy.

References

1. Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.: Surface reconstruction from unorganized points. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques, pp. 71–78 (1992)
2. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the 4th Eurographics Symposium on Geometry Processing, pp. 61–70 (2006)
3. Marton, Z.C., Rusu, R.B., Beetz, M.: On fast surface reconstruction methods for large and noisy point clouds. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 3218–3223 (2009)
4. Wang, J., Oliveira, M.M.: A hole-filling strategy for reconstruction of smooth surfaces in range images. In: Proceedings of the 16th Brazilian Symposium on Computer Graphics and Image Processing, pp. 11–18 (2003)
5. Lancaster, P., Salkauskas, K.: Surfaces generated by moving least squares methods. *Mathematics of Computation* 37(155), 141–158 (1981)
6. Schroeder, W.J., Zarge, J.A., Lorensen, W.E.: Decimation of triangle meshes. *ACM SIGGRAPH Computer Graphics* 26(2), 65–70 (1992)
7. Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. *Machine Learning* 3(2), 95–99 (1988)
8. Bevilacqua, V., Mastronardi, G., Piscopo, G.: Evolutionary approach to inverse planning in coplanar radiotherapy. *Image and Vision Computing* 25(2), 196–203
9. Michalewicz, Z., Nazhiyath, G.: Genocop III: A co-evolutionary algorithm for numerical optimization problems with nonlinear constraints. In: Proceedings of IEEE International Conference on Evolutionary Computation, vol. 2, pp. 647–651 (1995)
10. Sappa, A.D., Bevilacqua, V., Devy, M.: Improving a genetic algorithm segmentation by means of a fast edge detection technique. In: Proceedings of IEEE International Conference on Image Processing, ICIP, pp. 754–757 (2001)
11. Jones, M.W.: 3D distance from a point to a triangle. Technical Report, Dpt. of Computer Science, Univ. of Wales Swansea (1995)
12. Van der Bergen, G.: Efficient collision detection of complex deformable models using AABB trees. *Journal of Graphic Tools* 2(4), 1–13 (1997)
13. Bentley, J.L.: Multidimensional binary search trees used for associative searching. *Communications of the ACM* 18(9), 509–517 (1975)
14. Bevilacqua, V., Casorio, P., Mastronardi, G.: Extending Hough Transform to a Points’ Cloud for 3D-Face Nose-Tip Detection. In: Huang, D.-S., Wunsch II, D.C., Levine, D.S., Jo, K.-H. (eds.) *ICIC 2008*. LNCS (LNAI), vol. 5227, pp. 1200–1209. Springer, Heidelberg (2008)

Hexahedral Mesh Generation for Geometry with Multi-featured Constraints

Xing Dai¹, Han-Guo Cui^{1,*}, Li-Ping Zhang², Zheng-Min Li¹, and Fei-Zhang Wang¹

¹ College of Power Engineering, Naval University of Engineering, Wuhan, China
{dxxg1949, cuihanguo, lizhengm100, wangfeizhang}@163.com

² Wuhan Railway Vocational College of Technology, Wuhan, China
954516258@qq.com

Abstract. The accuracy and efficiency of finite element analysis is closely related to mesh quality. There is a lack of universal automatic meshing method currently. Existing hexahedral meshing methods suffer from poor meshing compatibility between different geometric constraints and poor meshing quality, especially for complex models with varieties of geometric features. To solve this problem, a novel hexahedral meshing method based on the mapping algorithm for multi-featured geometry is proposed. In this method, the model boundary is extracted and converted into regular shape, subdomain constraints generated by different geometric feature are projected to one face for centralization, the dividing number of boundary edges are comprehensive computed to make mapping mesh compatible between subdomains, and then dividing lines are generated by linking appropriate nodes on the subdomain boundary and the nodes on the face boundary with consideration about dividing result. Finally, high quality structured mesh is filled in the divided patches and then mapped back to the original model to finish hexahedral mesh generation. Examples demonstrate that the proposed method is reliable and robust, and can handle the hexahedral mesh generation of complex geometries with multi-features.

Keywords: finite element analysis, mapping method, hexahedral mesh, subdomain constraint.

1 Introduction

The finite element method can provide appropriate approximate numerical solution for problems of complex engineering and scientific, which is one of the most important methods in applied sciences and technology. The primary task of the finite element method is meshing, how to generate high quality mesh within the analysis area automatically and efficiently is the main "bottleneck" problem of the finite element method. In the meshing research these years, hexahedral mesh which has obvious advantages in computational accuracy, number of element, degree of anti-distortion becomes a hot research topic in recent years [1].

* Corresponding author.

Several main methods have been developed to generate hexahedral mesh for a given geometry: transform template methods [2], grid-based methods [3], advancing front methods [4] and mapping methods [5-7]. Compared to other methods, mapping method has advantages in mesh generation speed, mesh quality, mesh density controllability and has a wide application[5-7], this method divides the geometry into patches and then meshes each patch separately to finish the final hexahedral mesh. In its process, there are two big problems: how to divide the geometry effective and how to make patch mesh compatible, which make automation of the mapping method difficult. In order to overcome these two problems, this paper proposes a novel mapping method, which simplifies the complex model boundaries through space conversion, reduces the topological constraints through subdomain projection, and comprehensive computes dividing number of boundary edges to make patch mesh compatible. Several examples presented in the final illustrate the applicability of this method.

2 Hexahedral Mesh Generation

The hexahedral mesh method proposed contains five main steps: model boundary is extracted from STL format file; model boundary vertices are classified and converted to regular shape based on the vertices classification; dividing number of boundary edge is comprehensive computed; converted model topology is simplified and then dividing lines are generated; structured mesh (the nodes of mesh has clear and orderly position relationship) are filled in the patches divided and then mapped back to the original model.

2.1 Model Boundary Extraction

In order to achieve effective expression of the overall structure of the model, model boundary should be extracted from some format file. This paper chooses STL (stereolithography) format file which is common in most CAD (computer aided design) software. STL format file uses triangular facets to express model, which includes vertices coordinate, normal vector of triangular facets and has a good expression for geometries with changing boundary curvature. However triangular facets in the STL format is arranged disorderly and lack topology links between each other, the connecting nodes between adjacent triangular facets are stored repeatedly, which results in a lot of redundant information. Therefore, boundary information in the STL format file should be extracted and arranged reasonable before meshing. This paper first computes the angle between two adjacent triangular facets normal first, When the result angle is less than the threshold value (typically 20 degrees), two triangular can be considered to be coplanar, then judge the connecting property of each side of the triangle on the same surface, when selected side is not shared by the other triangle, the side selected is boundary edge, and then judge the connecting property of each node of the triangle, when selected node is shared by two or three boundary edge, the node selected is boundary node. After going through all the triangular facets in the STL format file according to the method described above, all the boundary information of the model can be extracted, as shown in Fig.1.

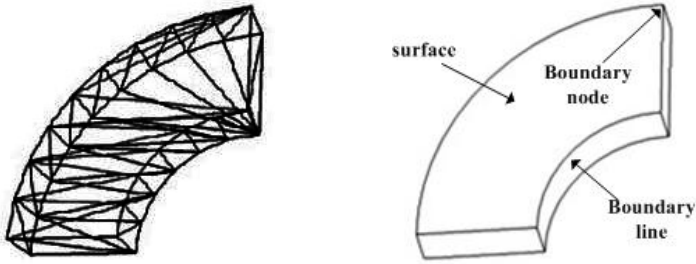


Fig. 1. Model boundary extraction from STL format file: (left) the STL model, (right) the boundary information extracted

2.2 Vertex and Edge Classification

Vertex classification is mainly based on the angle formed by its connecting edges in local surface [6], when the angle is close to $\pi/2$, the vertex is defined as "end"; when the angle is close to π , the vertex is defined as "side"; when the angle is close to $\pi*3/2$, the vertex is defined as "corner"; when the angle is close to $\pi*2$, the vertex is defined as "reversal". Edge classification is mainly based on the direction in local surface, set the direction of the boundary edge in the surface $f1$ with local 2D coordinate (i, j) in computational space counterclockwise, the edge in $+i$ direction (horizontal going from left to right) in the face $f1$ can be defined as i_{f1}^+ ; the edge in $-i$ direction (horizontal going from right to left) in the face $f1$ can be defined as i_{f1}^- ; the edge in $+j$ direction (vertical going upwards) in the face $f1$ can be defined as j_{f1}^+ ; the edge in $-j$ direction (vertical going downwards) in the face $f1$ can be defined as j_{f1}^- . Based on the vertex classification, adjust the angle formed by linking edges into standard angle (the standard angle of "end" is $\pi/2$; the standard angle of "side" is π ; the standard angle of "corner" is $\pi*3/2$; the standard angle of "reversal" is $\pi*2$) in computational space, then the complex models can be converted into regular shape as shown in Fig.2.

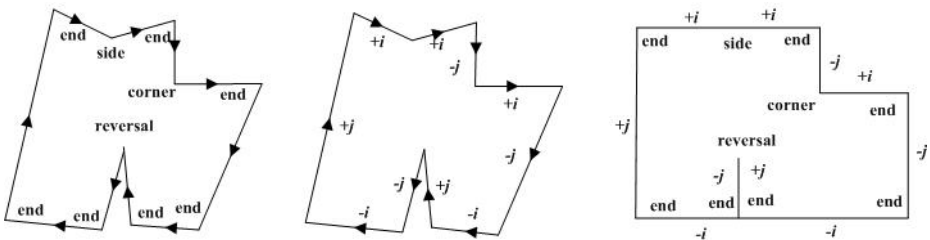


Fig. 2. Vertex and edge classification : (left) vertex classification; (center) edge classification; (right) model converted in computational space

2.3 Dividing Number of Edges Assignment

In order to obtain a structured mesh in the regular model in computational space, the dividing number of boundary lines in opposite should be equal as shown in Fig.3. Let N_k be the dividing number of the edge in a face, k is the serial number of the boundary edge, based on the vertex and edge classification described in Section 2.2, N_k should satisfy

$$\begin{cases} obj : \min \sum_k N_k \\ s.t : N_k > 0, N_k \in \mathbb{Z}^+ \\ N_k = L_k / L_{mesh} \\ f = \{f1, f2, f3 \dots\} \\ \sum_{k \in \{i^+\}} N_k = \sum_{k \in \{i^-\}} N_k, \sum_{k \in \{j^+\}} N_k = \sum_{k \in \{j^-\}} N_k \end{cases} \quad (1)$$

Where $\{+i\}$ represent the collection of the serial number of edges with $+i$ direction; $\{-i\}$ represent the collection of the serial number of edges with $-i$ direction; $\{+j\}$ represent the collection of the serial number of edges with $+j$ direction; $\{-j\}$ represent the collection of the serial number of edges with $-j$ direction, L_k is the actual length of the edge and L_{mesh} is the length of element assigned, f represents the collection of the model face.

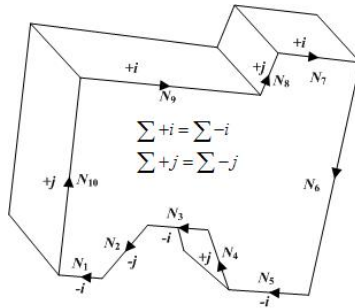


Fig. 3. Dividing number of edges computation

In order to ensure the dividing length of edge is close to L_{mesh} , set the target value (sum of the dividing number of edges) minimum. The equation described above can be solved by integer programming method[8].

2.4 Topology Simplification

After simplifying the model boundary in computational space, the subdomain boundary in each face should be concentrated in one face to ensure the mesh in this face can be compatible with subdomain boundary in other face, which is shown in Fig.4, Fig.4(left)present a model with three feature constraints, Fig.4(right) present three subdomains generated by features were projected in one face.

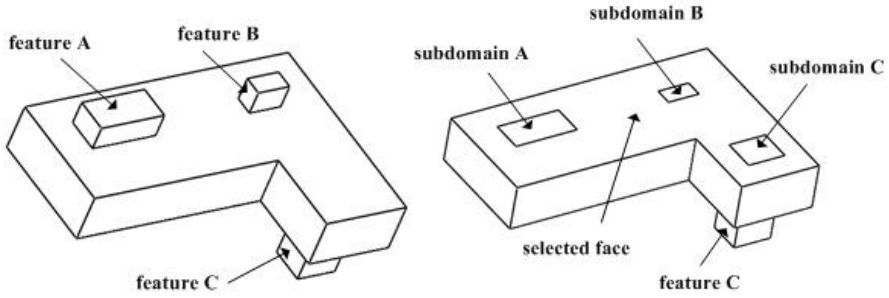


Fig. 4. Subdomains concentration:(left) original model;(right) subdomains projection

After subdomains concentration, new boundary lines should be generated through connecting vertexes on the subdomain boundary lines to the node on face boundary lines for topological simplification, all the nodes on lines are generated by assigning dividing number computed in section 2.3 to boundary lines, there are two factors should be considered when selecting new boundary lines, which is the length of the new boundary line and the angle formed by new boundary line and surface boundary. T_1 can be defined as

$$T_1 = a_1 \frac{\sum_{k=1}^2 (\theta_k - \pi / 2)}{2\pi} + a_2 \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{L_{\max}} \quad (2)$$

where a_1 and a_2 are the weight coefficient, θ_k is the angle formed by new boundary line and face boundary, (x_i, y_i) is the coordinate of the vertex on subdomain boundary lines, (x_j, y_j) is the coordinate of the node on face boundary lines, L_{\max} is the length of the longest new boundary lines. The first term of Eq.(2) ensures the dividing line internal angles are close to 90, the second term of Eq.(2) ensures the dividing lines is the shortest.

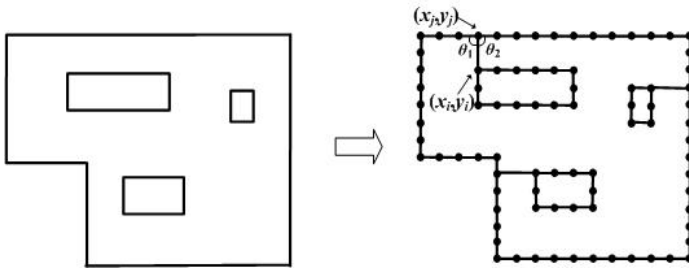


Fig. 5. Topology simplification

For a model with multi-subdoamins, the number of new boundary lines is equal to the number of subdomains, for example, the model in Fig.5 has three subdomains, and the number of new boundary lines is three.

2.5 Dividing Lines Generation

After topology simplification, the dividing lines are used to divide the model face, an optimal dividing lines dividing model face into two simpler regions R_1 and R_2 is determined via the parameter T_2 , T_2 is applied recursively to choose dividing line until the resulting subregion is simple enough. Let A_i be the area of region R_i , U_i be the circumference of the region R_i , l be the length of the dividing line, and then set $A=\min(A_1,A_2)$, $U=\min(U_1,U_2)$, parameter T_2 can be defined as

$$T_2 = b_1\left(\frac{l_2}{A}\right) + b_2\left(\frac{l}{U}\right) + b_3 \frac{\sum_{k=1}^2 (\theta_k - \pi / 2)}{2\pi} \tag{3}$$

Where b_1, b_2, b_3 are three weight coefficients, the first term is defined to ensure the resulting subregion is the smallest, the second term is defined to ensure the length of the dividing line is the shortest, the third term is defined to ensure dividing line internal angles are close to 90. With application of parameter T_2 , reasonable dividing lines can be generated to divide model face into simple convex regions as shown in Fig.6, the dividing number of dividing can be computed by following the dividing number of opposite line in the region is equal.

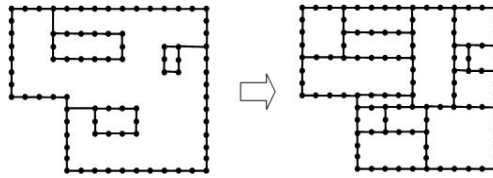


Fig. 6. Dividing lines generation

2.6 Structured Mesh Generation

After dividing the model in computational space, transfinite interpolation method[9] can be used to fill the structured mesh in the regions divided. Let $F(s, t)$ be a quadrilateral face, the computational space is defined as $[0,1] \times [0,1]$, and then the boundary of face can be defined as $F(s, 0), F(s, 1), F(0, t), F(1, t)$. As shown in Fig.7, mapping the mesh node from computational space to original model using transfinite interpolation method can be defined as

$$P(\xi, \eta) = (1-\eta)F(\xi, 0) + \eta F(\xi, 1) + (1-\xi)F(0, \eta) + \xi F(1, \eta) - (1-\xi)(1-\eta)F(0, 0) - (1-\xi)\eta F(0, 1) - \xi(1-\eta)F(1, 0) - \xi\eta F(1, 1) \tag{4}$$

Eq.(4) shows the arbitrary mesh node of surface can be decided by the nodes on the four boundary lines $F(s,0), F(s,1), F(0,t), F(1,t)$ and four vertexes of the boundary lines $F(0,0), F(1,0), F(1,1), F(0,1)$.

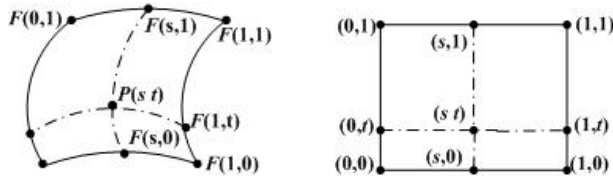


Fig. 7. Transfinite interpolation:(left) original model;(right) model in computational space. For a two-dimensional model, it is only needed to determine the coordinate of boundary node and vertex, quadrilateral mesh can be automatically generated using transfinite interpolation method, and hexahedral mesh can be generated by sweeping quadrilateral mesh finally.

3 Experimental Results

Fig.8a is a simple perforated model, one face of the model contains six vertexes defined as "end" and one vertex defined as "corner", the model is converted into a regular shape in computational space as shown in Fig.8b, Fig.8c shows the structured mesh is filled in the regular model in computational space, Fig.8d shows the hexahedral mesh mapping back to original model from computational space. Fig.9a is a complex model with two holes and curved boundary, based on the vertex and edge classification, the model is converted into regular shape in computational space as shown in Fig.9b, the structured mesh filled in computational mesh is shown as in Fig.9c, the final hexahedral mesh is shown as in Fig.9d, which is well correspondent with the structured mesh in the computational space without distortion. Fig.10(left) shows a regular model, the final hexahedral mesh generation using the method proposed shown in Fig.10(right) maintain a uniform shape without any distortion around hole. Jacobean value (a parameter reflecting the mesh quality, the more parameter value is close to 1, the better mesh quality is) of the final mesh as shown in Fig.8d, Fig.9d, Fig.10(right) are all above 0.8 and final mesh are all in orderliness, which proves that the method proposed is in high quality and efficient.

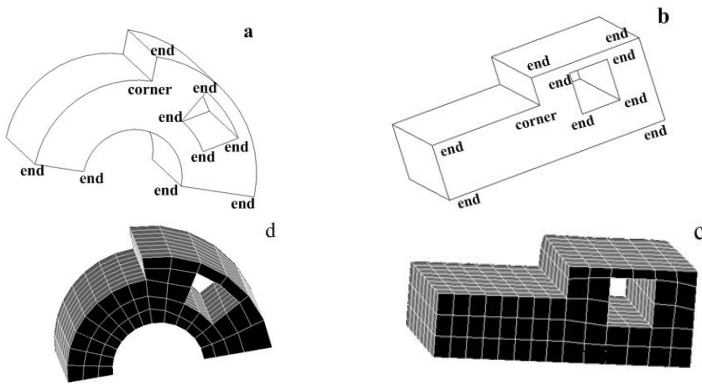


Fig. 8. Mesh for a perforated model:(a) original model;(b) converted model;(c) mesh in the converted model;(d) final mesh(Jacobean:0.82~1)

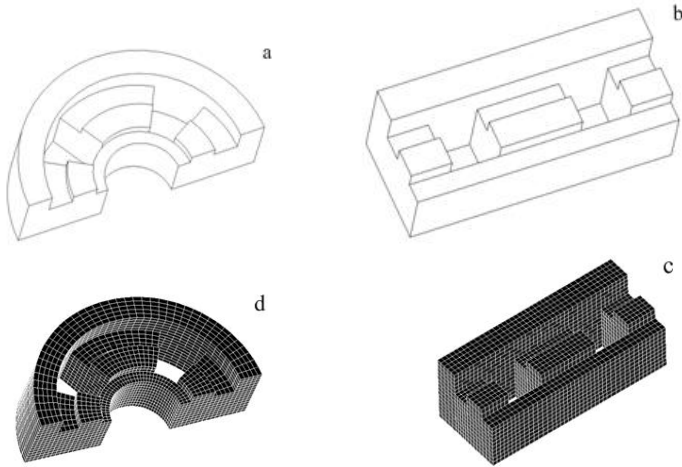


Fig. 9. Mesh for a perforated model:(a) original model;(b) converted model;(c) mesh in the converted model;(d) final mesh(Jacobean:0.85~1)

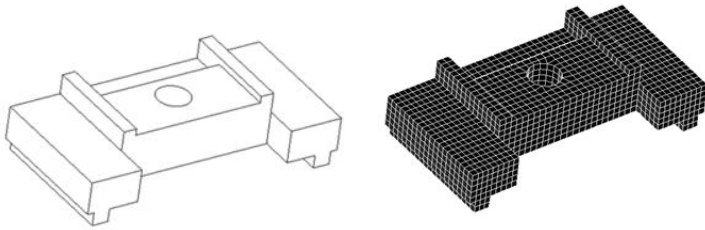


Fig. 10. Mesh for a model with regular boundary:(left) Original model;(right) final mesh(Jacobean:0.83~1)

4 Conclusions

In this work, we have present three contributions in order to improve the hexahedral mesh, the first one is simplifying the complex boundary through space conversion, the second one is concentrating the constraint formed by different feature on the face, which simplifies the model topology, the third one is filling the structured mesh in the model converted and then mapping the mesh back to original model, which improves the overall quality of the mesh. Experimental results show that the mesh generated by the method proposed does not need any smoothing operation and can be fully in order with high quality.

Acknowledgements. This work was partially sponsored by Natural Science Foundation of Hubei Province of China under grants 2012FB06904.

References

1. Schonning, A., Oommen, B., Ionescu, I., Conway, T.: Hexahedral Mesh Development of Free-Formed Geometry: the Human Femur Exemplified. *Computer Aided Design* 41(8), 566–572 (2009)
2. Guang, Z.Q., Shan, J.L., Gu, Y.X.: All Hexahedron Mesh Generation for 3D Solid Model Based on the extended Transform Templates. *Chinese Journal of Computational Mechanics* 22(1), 32–37 (2005)
3. Su, Y., Lee, K.H., Kumar, A.S.: Automatic Hexahedral Mesh Generation Using a New Grid-Based Method with Geometry and Mesh Transformation. *Computer Methods in Applied Mechanics and Engineering* 194(39), 4071–4096 (2005)
4. Kawamura, Y., Islam, M.S., Sumi, Y.: A Strategy of Automatic Hexahedral Mesh Generation by using an Improved Whisker-Weaving Method with a Surface Mesh Modification Procedure. *Engineering with Computers* 24(3), 215–229 (2008)
5. Ruiz-Gironés, E., Sarrate, J.: Generation of Structured Hexahedral Meshes in Volumes with Holes. *Finite Elements in Analysis and Design* 46(10), 792–804 (2010)
6. Ruiz-Gironés, E., Sarrate, J.: Generation of Structured Meshes in Multiply connected Surfaces Using Submapping. *Finite Elements in Analysis and Design* 41(2), 379–387 (2010)
7. Dai, X., Cui, H.G., Luo, X.: All Hexahedron Mesh Generation for Multi-Featured Geometry. *Chinese Journal of Computational Mechanics* 30(2), 224–230 (2013)
8. Whiteley, M., White, D., Benzley, S., Blacker, T.: Two and Three-Quarter Dimensional Meshing Facilitators. *Engineering with Computers* 12(3-4), 144–154 (1996)
9. Li, T.S., Armstrong, C.G., McKeag, R.M.: Quad Mesh Generation for K-Sided Faces and Hex Mesh Generation for Trivalent Polyhedra. *Finite Elements in Analysis and Design* 26(4), 279–301 (1997)

Computing Model of Individual Emotion in the Mass Incidents with Venting Anger

Fanliang Bu and Yiyi Wang

Department of Engineering of Security & Protection System,
Chinese People's Public Security University, Beijing, China
wangyiyi.0512@qq.com

Abstract. Based on emotional contagion theory and combined emotional contagion scale with modeling and simulating research methods then put forward a computing model of individual emotion in the mass incidents with venting anger. In the model, emotional contagion scale is used to measure emotional contagion sensitivity of the agents. The agents have parameters such as five basic emotions, explicit emotions, comprehensive emotions and so on. A large number of simulation results show that the model can simulate the changes of individual basic emotions, the group basic emotions and the group comprehensive emotions in the process of emotional contagion.

Keywords: the mass incidents with venting anger, emotional contagion scale, emotional contagion model.

1 Introduction

The mass incidents with venting anger (VAMI) happen in very occasional circumstances. They have no clear organizers and the vast majority of participants do not have a direct interest with the initial event. The participants vent their anger and can easily lead to mob violence then will cause extremely serious damage [1]. How to curb the multiple trend of such incidents has become a very difficult problem. In recent years, with the development of computer modeling and simulation technology, some scholars have studied to simulate mass violence incidents [2,3]. But in VAMI, we can not ignore the importance of the emotional theory during the study of modeling and simulating the mass incidents with venting anger. In this paper, we based on emotional contagion theory and combined emotional contagion scale with modeling and simulating research methods then put forward a computing model of individual emotion in VAMI.

2 Correlation Theory

2.1 Emotion and Emotional Characteristics of VAMI

The emotion that in VAMI have characteristics as follow:

Firstly, the emotion related to cognition. Emotion comes from a high degree of positive or negative evaluation. The organisms evaluate the significance of induced events or situations in a certain number of dimensions and the results decide the impending emotion. So, we considered evaluation is highly cognitive [4]. The VAMI participants produce emotions based on their own social attitudes, the past, the interests of the damage, etc. Secondly, the emotion is always changing. Relative to affection, emotions tend to exhibit its transience and mutability. In the VAMI forming process, participants experienced a changing process as emotional accumulated, emotional arousal, emotional contagion and emotional combustion and the temporarily changed emotion seriously affect the participant's behavioral tendencies. Finally, the interaction between individuals has great influence on the change of emotion. Under the condition of crowded gathered, psychological suggestion and behavior mimics can easily occur because of close contact. That will lead a great probability of occurrence of emotional contagion. A person's emotion infects another, another infects the others and the infected people will in return affect the first person. Finally, the whole crowd will combust their emotion. We main focus on third characteristics, and introduce the contagion theory of emotion firstly.

2.2 Emotional Contagion

Hatfield have finished some basic work about emotional contagion. They named this process "emotional contagion", and further defined it as a kind of tendency to automatically imitate and synchronized others' facial expressions, voice, posture, movements and behavior, and the results always make the emotions to get together and unified[5].

Barsade further reveals the influence of emotional contagion in groups. He studied that how the two factors which are emotional valence and emotional energy level to affect the process of emotional contagion. He pointed out that the reason why the group is not the set of independent individual is that the groups' internal have strong emotional interaction. This emotional interaction process is more complex and changeable than only between two independent individual. He also assumed that negative emotions more easily lead to emotional contagion, and when have the same emotion valence the high energy emotions cause more infections[6].

2.3 Emotional Contagion Scale

Doherty proposed using the emotion contagion scale to measure the differences of emotional contagion sensitivity in individual [7]. He designed a one-dimensional table with 15 terms for measuring persons' emotional contagion sensitivity. He set five basic emotions including love, happiness, fear, anger and sadness. The questionnaire using a five-point Likert scale which using never, rarely, sometime, often, always corresponding point 1 to 5. After questionnaire surveying, each emotion gets a score. Doherty verified the validity of emotional contagion scale.

3 Emotional Contagion Model

Depending on emotional contagion model, each agent contains 5 basic emotions including love, happiness, fear, anger and sadness. Together, they constitute the agents' emotional state. According to the type of emotion, the emotion is mapped to the two emotional impacts sets-Positive impacts (PI) and Negative impacts (NI). PI means the corresponding emotions will lead the individual does not participate in the event. NI means the corresponding emotions will lead the individual to participate in the event. In the 5 basic emotions, love, happiness and fear impact for PI, and the other impact for NI. When one agent impacted by other agent's emotion, the corresponding emotion variable values changing, the PI and NI together impacting the agent, and getting the agent's current emotion value. The initial emotional state affected by the individual experience, cognitive structure, purpose, and other factors. Also, each emotion correspond one ECS value. ECS used to indicate the possibility to be influenced by others of the agent. Each emotion's ECS value has different energy level, average and deviation[8].

Table 1. ECS value and emotional energy level

Emotion	Energy	ECS Mean (SD)
Happiness	Neutral	4.0 (0.6)
Love	High	4.0 (0.7)
Anger	High	3.5 (0.7)
Fear	High	3.1 (0.8)
Sadness	Low	3.4 (0.8)

At present, there are no studies have shown that low energy level emotion is less prone to lead emotional contagion. But it has been confirmed that the main emotions for low energy level of emotional agent are less susceptible to the emotions of others. ECS energy level and emotional relationship is fixed in the model. According to the characteristics of VAMI, set different energy level value for different emotions in the model.

In VAMI, the emotional contagion model shown in Figure 1.

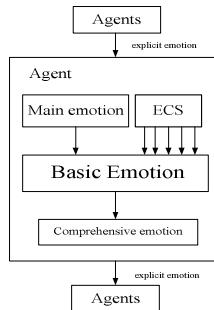


Fig. 1. Emotional contagion model

3.1 Agent Parameters

- Basic Emotion

The agent has five basic emotions including love, happiness, fear, anger and sadness. Each emotion corresponds to a parameter that used to record the value of every emotion.

- The ECS value of emotion

To simplify the model, we will have 1 to 5 of the ECS value corresponding to 0% to 100%. The emotional ECS value coincides with the normally distributed random value. The five emotions corresponding ECS energy level values for 0.1,0.05,0.1,0.1 and 0.02, respectively.

- The explicit emotion

The explicit emotion has the highest emotion value in the five basic emotions. In the process of emotional contagion, explicit emotions directly impact the other agents' emotions. According to the most of VAMI participants that with negative emotions, setting the individual initial explicit emotion that with a bigger probability than the other emotions is anger and sadness.

- Comprehensive emotion

The agents' comprehensive emotion shown the combined effects value that impacted by the positive and negative impacts. In VAMI, most of the agents have negative emotions. So, the agents' comprehensive emotion will go down because of love, happiness and fear. The higher comprehensive emotion the agents have, the higher negative emotion value they get.

3.2 Parameter Variation[9]

In the model, the emotion recipient defined as Agent r, the emotion sender defined as Agent s, Agents is both the recipient and the sender. r is the sight radius of Agent r, the Agents that within his sight can interact with him each other. The bigger and denser the crowd scale is, the larger the number of other Agents within sight will be. When only one Agent s, Agent r received emotional intensity is defined as:

$$E_r^* = ECS * energy \quad (1)$$

Among them, ECS is the ECS value and energy is the energy level value that corresponding to the explicit emotion that received by Agent r. When Agent r and Agent s have the same explicit emotion, adjusting the emotional intensity as:

$$emotion(t+1) = emotion(t) + ECS * energy \quad (2)$$

The other emotions of Agent r is:

$$emotion(t + 1) = emotion(t) - ECS * energy \tag{3}$$

So, all within sight Agent s' emotional intensity received by Agent r is:

$$E_r = \sum_s E_r^* \tag{4}$$

If the explicit emotion of Agent s is love, happiness and fear, there would be a positive impact effect on Agent r, or there would be a negative impact. The positive impact received by Agent r is represented as:

$$P = e_r e_{sp} \tag{5}$$

e_{sp} represents the weighted sum of positive emotional intensity sent by the sender.

$$e_{sp} = \sum_{sp} \frac{ECS * energy}{\sum_s ECS * energy} currentemotion_s \tag{6}$$

Similarly, the positive impact received by Agent r is as follows:

$$N = e_r e_{sn} \tag{7}$$

e_{sn} represents the weighted sum of negative emotional intensity sent by the sender.

$$e_{sn} = \sum_{sn} \frac{ECS * energy}{\sum_s ECS * energy} currentemotion_s \tag{8}$$

So, the comprehensive emotion of Agent r is:

$$e_r(t+1) = e_r(t) + E_r \left[\sum emotion(t) * N - \sum emotion(t) * P \right] \tag{9}$$

$emotion$ is the Agent r emotion value corresponding to the positive impact and negative impact.

4 Simulation Experiments and Results Analysis

4.1 Effects of Crowded Scale on Crowded Basic Emotions

Setting the crowded scale is 10, 30, 50 and 100, observing the changes of the average value of crowded basic emotion. The simulation results is shown as Figure 2.

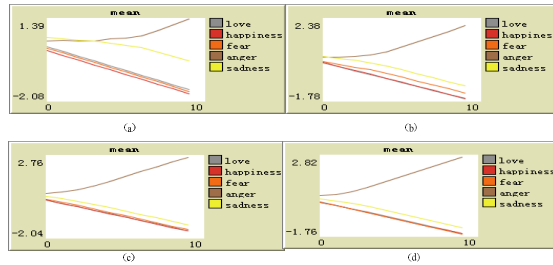


Fig. 2. The crowd basic emotion changes under the different crowd scale condition

The simulation results show that although the crowd have different scales, they have the same explicit emotion after impacted by anger. When the scale of crowd is 10, shown as Figure 2(a), the explicit emotion value of anger is 1.39. When the scale is 30, shown as Figure 2(b), the explicit emotion value of anger is 2.69. When the scale is 100, shown as Figure 2(d), the explicit emotion value of anger is 2.81. This shows that with the increase of the scale of the crowd, the explicit emotion value of anger also becomes bigger. But when the scale is big enough, the amplitude of increasing of emotion value goes down. In the VAMI, when the emotions of crowd are dissatisfaction and anger, the bigger the scale of crowd is, the stronger the negative emotions. The simulation results are consistent with the real situation.

4.2 Effects of Initial Explicit Emotional on Individual Emotion

Setting the scale of the crowd is 20, randomly selected individual from the crowd with different initial explicit emotions, observing their basic emotion changes, the simulation results shown as Figure 3:

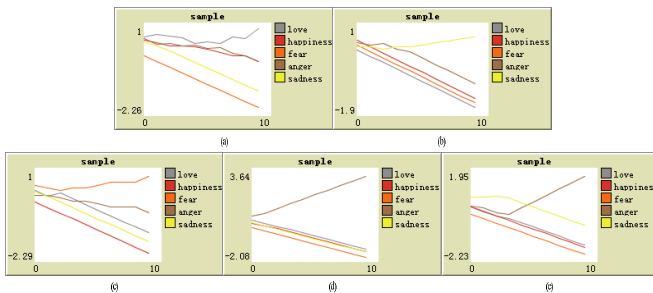


Fig. 3. Basic emotion changes of individual with different initial explicit emotion

Shown as Figure 3(a), the initial explicit emotion of sample is love, this is a positive emotion with a high energy level. The agent whose emotion is love is difficult to be changed by the others' emotion. In the VAMI, this kind of agents can be treated as the individual with satisfaction emotion. Their participation in the event probability is very low. Shown as Figure 3(b), the initial explicit emotion of sample is happiness, this is a positive emotion with a general strength of the energy level. The agent whose emotion

is happiness may be changed because of the others' emotion. In the VAMI, this kind of agents can be treated as spectators. They may both become dissatisfied and participate in the event, and may also choose to leave. Shown as Figure 3(c), the initial explicit emotion of sample is fear, this is emotion with a high energy level. In the VAMI, this kind of agents can be treated as the individual with fear. They may choose to leave as quickly as possible. Shown as Figure 3(c), the initial explicit emotion of sample is fear, this is a negative emotion. In the VAMI, this kind of agents can be treated as the individual with extreme dissatisfaction. They become more dissatisfied because of emotion contagion. So they may out of their senses and do aggressive behavior.

Shown as Figure 3(e), the initial explicit emotion of sample individual is sadness. This is a negative emotion with low energy level. After emotional contagion, the initial explicit emotion of sample individuals changed, it turned to be anger. In the VAMI, this kind of agent represent the individuals that with dissatisfaction but insensitive. They would participate in the event if they affected by militants.

As the crowd is consisted of by different individuals, we should take appropriate ways to overawe the cowed when we deal with the VAMI. Persuading and educating the easily transformed ones, focusing on the disposal of militants. Crushing them one by one, collapsing the crowd at last.

4.3 Effects of Explicit Emotion of Crowded on Comprehensive Emotion

Setting the scale of crowd is 20, observing the changes of average comprehensive emotion when the crowd has different explicit emotions. The simulation result is shown as Figure 4.

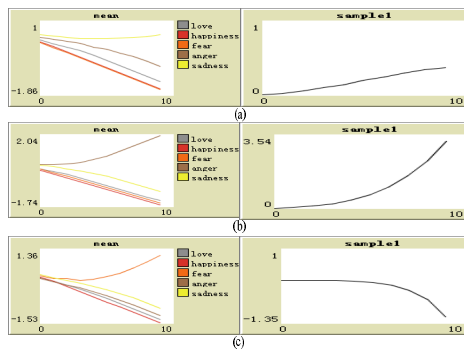


Fig. 4. The changes of comprehensive emotion when the crowd has different explicit emotions

Shown as Figure 4(a), the explicit emotion of crowd is sadness. The comprehensive emotion increases because of the impact of negative emotion. But it tent to be saturated. In Figure 4(b), the explicit emotion is anger. The crowd is impacted by high energy negative emotion. The average comprehensive emotion increases quickly. This is thus clear that in the VAMI, if the individual dissatisfaction degree is small, after emotional contagion, they still can be controlled easily because the negative emotion strength is small. If the individual is extremely dissatisfied, after emotional contagion, the individual dissatisfied emotion would increases quickly until the crowd lost of

control. In Figure 4(c), the explicit emotion is fear, the average comprehensive emotion reduce. The individuals that have fear emotion are afraid of risk. In the VAMI, this kind of individual will leave away. In the VAMI, the dissatisfaction degree decides the development direction and the degree of control of the event. So if we can take effective measures to release the dissatisfied emotion, thus can we largely control the development direction of VAMI.

5 Conclusion

This paper is based on emotional contagion theory, combined emotional contagion scale with modeling and simulating research methods, then put forward a computing model of individual emotion in the VAMI. By analysis the tests and simulation results we can reach a conclusion that, in the VAMI, although the scale of the crowd is different, they have the same explicit emotion after emotional contagion, that is anger. With the increase of crowd scale, the explicit anger emotion value becomes larger. But when the scale is big enough, the amplitude of increasing of anger emotion value goes down. In the modeling process we ignore the influence of many other factors on individual emotions, such as the weather, gender, individual education etc. In the future work, we will further.

References

1. Zhang, J.L.: Mass incidents at all times and in all over the world, pp. 133–134. Intellectual Property Publishing House, Beijing (2011)
2. Bu, F.L., Feng, P.Y.: Analysis of Agent-based non-organization and direct interest collective event. In: Tseng, A.A., Goertzel, B. (eds.) *Automatic Control and Artificial Intelligence: 3*, pp. 1762–1767. IEEE, Xiamen (2012)
3. Bu, F.L., Sun, J.: Particle Swarm Optimization-based Simulation and Modeling System for Mass Violence Events. In: Tseng, A.A., Goertzel, B. (eds.) *Automatic Control and Artificial Intelligence: 3*, pp. 2093–2097. IEEE (2012)
4. Du, J.: Emotional contagion and face research in service remedy, vol. 53. Nankai University Press (2010)
5. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Primitive emotional contagion. *Emotion and Social Behavior* 14, 151–177 (1992)
6. Barsade, S.G.: Emotional Contagion and Its Influence on Group Behavior. *Administrative Science Quarterly* 47(1), 644–675 (2002)
7. Doherty, R.W.: A Measure of Individual Differences. *Journal of Nonverbal Behavior* 21, 131–154 (1997)
8. Lundqvist, L.O.: The relationship between the Biosocial Model of Personality and susceptibility to emotional contagion: A structural equation modeling approach. *Personality and Individual Differences* 45, 89–95 (2008)
9. Bosse, T., Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: A Multi-agent Model for Emotion Contagion Spirals Integrated within a Supporting Ambient Agent Model. In: Yang, J.-J., Yokoo, M., Ito, T., Jin, Z., Scerri, P. (eds.) *PRIMA 2009. LNCS*, vol. 5925, pp. 48–67. Springer, Heidelberg (2009)

A SVM-Based System for Predicting Protein-Protein Interactions Using a Novel Representation of Protein Sequences

Zhuhong You¹, Zhong Ming¹, Ben Niu¹, Suping Deng², and Zexuan Zhu¹

¹ College of Computer Science and Software Engineering, Shenzhen University
Shenzhen, Guangdong 518060, China

² Department of Computer Science and Technology, Tongji University
Shanghai 201804, P.R. China
zhyou@szu.edu.cn

Abstract. Protein-protein interactions (PPIs) are crucial for almost all cellular processes, including metabolic cycles, DNA transcription and replication, and signaling cascades. However, the experimental methods for identifying PPIs are both time-consuming and expensive. Therefore, it is important to develop computational approaches for predicting PPIs. In this article, a sequence-based method is developed by combining a novel feature representation using binary coding and Support Vector Machine (SVM). The binary-coding-based descriptors account for the interactions between residues a certain distance apart in the protein sequence, thus this method adequately takes the neighboring effect into account and mine interaction information from the continuous and discontinuous amino acids segments at the same time. When performed on the PPI data of *Saccharomyces cerevisiae*, the proposed method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%. Extensive experiments are performed to compare our method with the existing sequence-based method. Achieved results show that the proposed approach is very promising for predicting PPI, so it can be a useful supplementary tool for future proteomics studies.

Keywords: protein-protein interaction, binary coding, support vector machine, protein sequence, local descriptor.

1 Introduction

Proteins are crucial for almost all of functions in the cell, including metabolic cycles, DNA transcription and replication, and signaling cascades. Usually, proteins rarely perform their functions alone; instead they cooperate with other proteins by forming a huge network of protein-protein interactions (PPIs). PPIs are responsible for the majority of cellular functions. In the past decades, many innovative techniques for detecting PPIs have been developed [1-3]. Due to the progress in large-scale experimental technologies such as yeast two-hybrid (Y2H) screens [2, 4], tandem affinity purification (TAP) [1], mass spectrometric protein complex identification

(MS-PCI) [3] and other high-throughput biological techniques for PPIs detection, a large amount of PPIs data for different species has been accumulated [1-6]. However, the experimental methods are costly and time consuming, therefore current PPI pairs obtained from experiments only covers a small fraction of the complete PPI networks [7, 8]. In addition, large-scale experimental methods usually suffer from high rates of both false positive and false negative predictions [9-12]. Hence, it is of great practical significance to develop the reliable computational methods to facilitate the identification of PPIs [7].

A number of computational methods have been proposed for the prediction of PPIs based on different data types, including phylogenetic profiles, gene neighborhood, gene fusion, and sequence conservation between interacting proteins, literature mining knowledge. There are also methods that combine interaction information from several different data sources [13]. However, these methods cannot be implemented if such pre-knowledge about the proteins is not available. Recently, a couple of methods which derive information directly from amino acid sequence are of particular interest [14, 15]. Many researchers have engaged in the development of sequences-based method for discovering new PPIs, and the experiment results showed that the information of amino acid sequences alone is sufficient to predict PPIs. Among them, one of the excellent works is a SVM-based method developed by Shen et al [15]. In the study, the 20 amino acids were clustered into seven classes according to their dipoles and volumes of the side chains, and then the conjoint triad method abstracts the features of protein pairs based on the classification of amino acids. When applied to predict human PPIs, this method yields a high prediction accuracy of 83.9%. Because the conjoint triad method cannot takes neighboring effect into account and the interactions usually occur in the discontinuous amino acids segments in the sequence, on the other work Guo et al. developed a method based on SVM and auto covariance to extract the interactions information in the discontinuous amino acids segments in the sequence [7]. Their method yielded a prediction accuracy of 86.55%, when applied to predicting *saccharomyces cerevisiae* PPIs. In our previous works, we also obtained good prediction performance by using autocorrelation descriptors and correlation coefficient, respectively [16,17].

In this study, we report a new sequence-based method for the prediction of interacting protein pairs using SVM combined with binary coding. More specifically, we first represent each protein sequence as a vector by utilizing a binary-coding-based representation of protein sequence which provides us with a chance to mine interaction information from the continuous and discontinuous amino acids segments at the same time [18]. The effectiveness of binary-coding-based descriptors depends largely on the correct selection of amino acid grouping [18]. By grouping amino acids into a reduced alphabet, we can create a more accurate protein sequence representation. Here, we adopted the amino acids grouping according to the successful use of classification in [14]. Then we characterize a protein pair in different feature vectors by coding the vectors of two proteins in this protein pair. Finally, an SVM model is constructed using these feature vectors of the protein pair as input. To evaluate the performance, the proposed method was applied to *Saccharomyces cerevisiae* data. The experiment results show that our method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%.

2 Materials and Methodology

In this paper, we have presented a new approach to predict PPIs using support vector machine (SVM) from protein sequences. Our method for predicting the PPIs depends on three steps: (1) Generation of the PPI dataset; (2) Feature vector extraction; (3) Classification using SVM.

2.1 Generation of the Data Set

We evaluated the proposed method with the data from yeast used in the study of Guo et al. [7]. The PPI dataset was collected from *Saccharomyces cerevisiae* core subset of Database of Interacting Proteins (DIP). After the redundant protein pairs which contain a protein with fewer than 50 residues or have $\geq 40\%$ sequence identity were removed, the remaining 5594 protein pairs comprise the final positive dataset. The 5594 non-interacting protein pairs were generated from pairs of proteins whose sub-cellular localizations are different. The whole dataset consists of 11188 protein pairs, where half are from the positive dataset and half are from the negative dataset. Note that we have used exactly the same non-redundant dataset as used in Guo et al. [7]. Four-fifths of the protein pairs from the positive and negative dataset were respectively randomly selected as the training dataset and the remaining one-fifths were used as the test dataset.

2.2 Feature Vector Extraction

To use machine learning methods to predict PPIs from protein sequences, one of the most important computational challenges is to extract feature vectors from protein sequences in which the important information content of proteins is fully encoded. In this section, we adopt a novel sequence representation model by using binary coding based descriptors.

There are three types of local descriptors used in the aforementioned studies: Composition, Transition and Distribution, which are computed based on the variation of occurrence of functional groups of amino acids within the primary sequence of the protein. In this study, the 20 amino acids were firstly clustered into seven functional groups based on the dipoles and volumes of the side chains. The functional groups used were: Cluster_1 (amino acids A,G,V), Cluster_2 (amino acids C), Cluster_3 (amino acids D,E), Cluster_4 (amino acids F,I,L,P), Cluster_5 (amino acids H,N,Q,W), Cluster_6 (amino acids K,R) and Cluster_7 (amino acids M,S,T,Y). In total there would be 63 features (7 composition, 21 transition, 35 distribution) if they were computed from the whole amino acid sequence.

In order to extract the interaction information of protein sequences, we split the protein sequences into fifteen different regions of varying length and composition to describe multiple overlapping continuous and discontinuous interaction patterns within a protein sequence. We first divided the entire protein sequence into four equal length regions (A-D). Then a novel binary-coding-based method was adopted to construct a couple of continuous and discontinuous regions on the basis of above

partition (A-D). Here the continuous regions are composed of residues which are local in the polypeptide sequence, while discontinuous regions consist of residues from different parts of the sequence, brought together by the folding of the protein to its native structure.

More specifically, a protein sequence was encoded as the combination of 4-bit binary digits (0 or 1), which means we need fifteen different combinations (0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111). It should be noticed that here 0 or 1 denote one of the four equal length region A-D is excluded or included in constructing the continuous or discontinuous regions respectively. For example, 0011 denotes a continuous region constructed by C and D (the final 50% of the sequence). Similarly, 1011 represents a discontinuous region constructed by A, C and D (the first 25% and the final 50% of the sequence). These regions are illustrated in Figure 1. For each region the 63 local descriptors are extracted, resulting in a $63 \times 15 = 945$ feature vector. Then the PPI pair is characterized by concatenating the two vector spaces of two individual proteins. Thus, an 1890-dimensional vector has been constructed to represent each protein pair and used as a feature vector for input into SVM classifier.

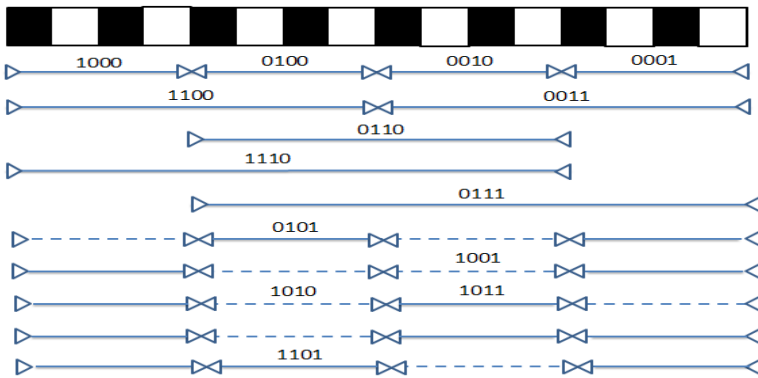


Fig. 1. The Schematic diagram for constructing fifteen descriptor regions for a hypothetical protein sequence

2.3 Support Vector Machine

Support Vector Machine (SVM) is a classification and regression paradigm first developed by Vapnik [19]. It has attracted much research attention in these years due to its improved generalization performance over other techniques in many real world applications including bioinformatics. The SVM originated from the idea of the structural risk minimization theory [19]. The main difference between this technique and many other conventional classification techniques including neural networks is that it minimizes the structural risk instead of the empirical risk. The principle is based on the fact that minimizing an upper bound on the generalization error rather than minimizing the training error is expected to perform better. SVM training always

seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book [20].

The basic idea of utilizing SVM model for classification can be stated briefly as follows. Firstly, map the original data X into a feature space F with high dimensionality through a linear or non-linear mapping function, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyperplane which separates the data into two classes.

Given a training dataset of instance-label pairs $\{x_i, y_i\}, i=1,2,\dots,N$ with input data $x_i \in R^n$ and labeled output data $y_i \in \{+1, -1\}$. The classification decision function implemented by SVM is represented in the following equation:

$$y(x) = \text{sign} \left[\sum_{i=1}^N y_i \alpha_i \cdot K(x, x_i) + b \right] \tag{1}$$

where the coefficients α_i are obtained by solving the following convex Quadratic Programming (QP) problem:

$$\text{Maximize} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \cdot y_i y_j \cdot K(x_i, x_j) \tag{2}$$

$$\text{Subject to} \quad 0 \leq \alpha_i \leq C \tag{3}$$

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad i = 1, 2, \dots, N. \tag{4}$$

In the equation (3), C is a regularization parameter which controls the tradeoff between margin and misclassification error. These x_j are called Support Vectors only if the corresponding $\alpha_j > 0$.

In this work, Radial Basis Functions (RBF) kernel, $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, is applied, which has better boundary response and most high-dimensional data sets can be approximated by Gaussian like distributions. In the experiment the well-known software LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) was employed to do classification.

3 Experiments and Results

3.1 Evaluation Measures

To evaluate the prediction performance of the proposed method, Sensitivity (Sens), Precision (PE), Matthews's correlation coefficient (MCC), and overall accuracy (Accu.) were calculated. The definitions of these measures are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$PE = \frac{TP}{TP + FP} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (8)$$

where true positive (TP) is the number of true PPIs that are predicted correctly; false negative (FN) is the number of true PPIs that are predicted to be non-interacting pairs; false positive (FP) is the number of true non-interacting pairs that are predicted to be PPIs, and true negative (TN) is the number of true non-interacting pairs that are predicted correctly. MCC denotes Mathew's correlation coefficient.

3.2 Prediction Performance of Proposed Model

We evaluated the performance of the proposed approach using the DIP PPIs data as investigated in Guo *et al.* [7]. In order to reduce the bias of training and testing data, a 5-fold cross-validation technique is adopted. More specifically, the dataset is divided into 5 subsets, and the holdout method is reiterated 5 times. Each time four of the five subsets are put together as the training dataset, and the other one subset is utilized for testing the model. Thus five models were generated for the five sets of data. The prediction results of SVM prediction models with proposed representation of protein sequence are shown in Table 1.

Table 1. The prediction result of the test dataset using proposed method

Classification Model	Testing set	Sens. (%)	Prec. (%)	Accu. (%)	MCC (%)
Proposed Method	1	86.60	86.99	86.72	73.45
	2	86.56	87.09	86.54	73.08
	3	87.60	87.21	87.53	75.05
	4	86.65	85.96	86.28	72.56
	5	87.55	87.23	87.59	75.17
	Average	86.99 ± 0.53	86.90 ± 0.53	86.93 ± 0.59	73.86 ± 1.18
Davies' Method	1	77.57	82.79	80.8672	61.84
	2	66.19	84.77	76.9781	55.39
	3	71.70	84.82	79.7944	60.17
	4	73.88	84.63	80.2414	60.97
	5	69.77	84.11	77.9018	56.81
	Average	71.82 ± 4.28	84.22 ± 0.85	79.16 ± 1.65	59.04 ± 2.79

It can be observed from Table 1 that for all five models the precisions are $\geq 85.96\%$, the sensitivities are $\geq 86.56\%$, and the prediction accuracies are $\geq 86.28\%$. On average, proposed method yields a PPI prediction model with an accuracy of $86.93 \pm 0.59\%$. To better investigate the practical prediction performance of proposed method, we also calculated the MCC value. From table 1, we can see that proposed method gives good prediction performance with an average MCC value of 73.86%. Further, it can also be seen in the experiments that the standard deviation of sensitivity, precision, accuracy and MCC are as low as 0.53%, 0.53%, 0.59% and 1.18% respectively. From the results, it can be concluded that proposed method is an accurate and robust method for the prediction of PPIs.

Many other sequence-based methods have been used for predicting of PPIs. In order to evaluate the prediction ability of the SVM prediction model using binary coding, extensive experiments are performed to compare our method with state-of-the-art techniques Davies' work [21]. Table 1 gives the average prediction results of 5-fold cross-validation over there two methods. From Table 1, we can see that the model based on Davies' work gives poor results with the average sensitivity, precision and accuracy of 71.82%, 84.22% and 79.16%, respectively. The results illustrate that our method outperforms other sequence-based methods such as Davies' method. All the analysis shows that our model is an accurate and fast method for the prediction of PPIs.

4 Discussion and Conclusions

With the large amount of protein sequences information provided by genome sequencing project, there is a growing demand for developing advanced computational methods for predicting potential PPIs using sequence information alone. In this study, we proposed a novel sequence-based approach for PPIs prediction using SVM combined with a binary-coding-based method. The binary-coding-based method was implemented to extract sequence information of proteins, and then an SVM algorithm was employed to construct the prediction model. The proposed representation of protein sequence descriptor account for the interactions between residues in both continuous and discontinuous regions of a protein sequence, so this method enables us to draw more PPI information from the protein sequence. When performed on the PPI data of *S.cerevisiae*, the method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%. Given the complex nature of PPIs, the performance of our method is promising and it can be a helpful supplementary for PPIs prediction.

Acknowledgments. This work is supported in part by the National Science Foundation of China, under Grants 61102119, 61170077, 71001072, 61272339, 61272333, 61001185, in part by the China Postdoctoral Science Foundation, under Grant 2012M520929.

References

- Gavin, A.C., Bosche, M., Krause, R., Grandi, P.: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868), 141–147 (2002)
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98(8), 4569–4574 (2001)
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L.: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415(6868), 180–183 (2002)
- Krogan, N.J., Cagney, G., Yu, H.Y., Zhong, G.Q.: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440(7084), 637–643 (2006)
- Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R.: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770), 623–627 (2000)
- Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y.: A protein interaction map of *Drosophila melanogaster*. *Science* 302(5651), 1727–1736 (2003)
- Guo, Y., Yu, L., Wen, Z., Li, M.: Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Research* 36(9), 3025–3030 (2008)
- You, Z.H., Yin, Z., Han, K., Huang, D.S., Zhou, X.: A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *Bmc Bioinformatics* 11 (2010)
- You, Z.H., Lei, Y.K., Gui, J., Huang, D.S., Zhou, X.: Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* 26(21), 2744–2751 (2010)
- Xia, J.F., You, Z.H., Wu, M., Wang, S.L., Zhao, X.M.: Improved method for predicting pi-turns in proteins using a two-stage classifier. *Protein and Peptide Letters* 17(9), 1117–1122 (2010)
- Lei, Y.K., You, Z.H., Ji, Z., Zhu, L., Huang, D.S.: Assessing and predicting protein interactions by combining manifold embedding with multiple information integration. *Bmc Bioinformatics* 13 (2012)
- You, Z.-H., Li, L., Yu, H., Chen, S., Wang, S.-L.: Increasing reliability of protein interactome by combining heterogeneous data sources with weighted network topological metrics. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) *ICIC 2010*. LNCS, vol. 6215, pp. 657–663. Springer, Heidelberg (2010)
- Qi, Y.J., Seetharaman, J.K., Joseph, Z.B.: Random forest similarity for protein-protein interaction prediction from multiple sources. In: *Pac. Symp. Biocomput.*, pp. 531–542 (2005)
- Yang, L., Xia, J.F., Gui, J.: Prediction of Protein-Protein Interactions from protein sequence using local descriptors. *Protein and Peptide Letters* 17(9), 1085–1090 (2010)
- Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H.: Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America* 104(11), 4337–4341 (2007)
- Shi, M.G., Xia, J.F., Li, X.L., Huang, D.S.: Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. *Amino Acids* 38(3), 891–899 (2010)

17. Xia, J.F., Han, K., Huang, D.S.: Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein and Peptide Letters* 17(1), 137–145 (2010)
18. Tong, J.C., Tammi, M.T.: Prediction of protein allergenicity using local description of amino acid sequence. *Frontiers in Bioscience* 13, 6072–6078 (2008)
19. Herrera, L.J.: Recursive prediction for long term time series forecasting using advanced models. *Neurocomputing* 70(16), 2870–2880 (2007)
20. Cortes, C., Vapnik, V.: Support vector network. *Machine Learning* (1995)
21. Davies, M.N., Secker, A., Freitas, A.A., Clark, E., Timmis, J., Flower, D.R.: Optimizing amino acid groupings for GPCR classification. *Bioinformatics* 24(18), 1980–1986 (2008)

Exploiting SAaaS in Smart City Scenarios

Salvatore Distefano¹, Giovanni Merlino^{2,3}, and Antonio Puliafito²

¹Dip. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy
salvatore.distefano@polimi.it

²Dip. di Ingegneria, Università di Messina, 98166 Messina, Italy
{gmerlino, apuliafito}@unime.it

³Dip. di Ingegneria (DIEEI), Università di Catania, 95125 Catania, Italy
giovanni.merlino@dieei.unict.it

Abstract. Most of the current shortcomings in relation to Clouds made up of sensing resources can be addressed, as detailed in past work from the authors, following a Sensing and Actuation as a service (SAaaS) approach, i.e. enabling an Infrastructure-oriented (IaaS-like) provisioning model for sensors and actuators. The aforementioned approach matches the technological requirements and constraints springing around certain application domains belonging to the Future Internet research area. One of the most prominent scenarios is computer-assisted data treatment and automation of urban areas and public facilities, going under the umbrella term of a so-called Smart City.

Aim of this paper is to provide specific processes, rules and guidelines for the adoption of the SAaaS paradigm in IT infrastructure powering Smart Cities. Such artifacts are for assessing the feasibility of deploying Smart City applications over SAaaS, moreover the design of some specific use cases is presented and discussed.

Keywords: Cloud, sensors and actuators, Smart Cities, mobile crowdsensing.

1 Introduction

According to current ICT trends (Internet of Things, Future Internet) and estimates soon there are going to be myriads of devices dispersed and meshed into Internet-wide networks over geographic distances, calling for infrastructure management systems and facilities at previously never attempted before reach and scale. Yet up till now sensing resources to be involved in a Cloud have been envisioned only as data collection outlets. The authors took a different trajectory in [1] to tackle transducers in a Cloud environment: sensing and actuation devices should be handled along the same lines as computing and storage abstractions in traditional Clouds, i.e. virtualized and multiplexed over (scarce) hardware resources. This brings about a so-called “Sensing and Actuation as a Service” (SAaaS), a provisioning model where (virtual) sensors and actuators are to be exposed as Infrastructure and provided as a (Web) service to customers.

In a SAaaS-enabled scenario nodes can be either fixed or roaming, contributed on a voluntary basis by owners, thus joining and leaving the system at will, unpredictably.

Our device-driven approach implies giving customers full control over virtual resources under rental. Among interesting applications for SAaaS, Smart City-related ones could play a remarkable role. Sensing and network technologies are core elements for designing and building new Smart City facilities, as well as for monitoring and maintaining the existing ones. Moreover, production, business and social processes could be profiled to optimize, e.g., yields and reduce energy consumption and environmental impact.

The Smart City application domain is reaching a prominent position among innovation trends. Future Internet and Internet of things place Smart Cities as a key concept for future technological developments [2]. The transformational changes towards smarter cities will require innovation in planning, management, and operations. Several ongoing projects illustrate the opportunities and challenges of this evolutionary step [3] in diverse application fields, e.g., networking, decision support-systems, power grids, energy-aware platforms, service oriented architecture, highlighting the need to equip the cities of the future with different types of urban sensors. Plenty of applications can be envisioned for a Smart City, from traffic monitoring to energy management, from e-health to e-government, from crowd management to emergency management and disaster recovery, etc. envisioning and unlocking new forms, schemes and mechanisms of (micro/crowd-) business activities and funding. This all-encompassing and very ambitious scenario calls for adequate ICT technologies. In particular, infrastructural solutions for managing underlying physical sensing and actuation resources are required. This paper focuses on ways to deal with Smart City challenges through the SAaaS infrastructure. Selected guidelines, rules and algorithms are specified in the following, describing how to setup both the SAaaS infrastructure and the application, and also how to deploy the latter into the SAaaS infrastructure thus obtained.

Several works deal with issues related to Smart Cities. An interesting investigation about similarities and differences between “smart” and “digital” cities is provided by [4]. Starting from this, the authors of [5] propose a platform for managing urban services that include convenience, health, safety, and comfort. Also Cloud computing infrastructure recently found useful application in the context of Smart Cities [6-7]. Even if a lot of applications in the Smart City scenario have been proposed so far, there is a lack of common initiatives and strategies to address infrastructural issues. Indeed, the infrastructure adopted in such applications implements ad-hoc solutions. SAaaS aims at filling this gap between Smart City applications and the underlying infrastructure.

The remainder of the paper is thus organized as follows. Section 2 describes and investigates the Smart City scenario from the infrastructural perspective, contextualizing the SAaaS and providing a high-level setup process overview. Then, Section 3 details the first setup phase, the SAaaS provider and infrastructure setup, also briefly discussing the SAaaS approach and related modular architecture, while Section 4 deals with the application setup and deployment, detailing different Smart City use cases. Section 5 closes the paper with some final remarks and discussion.

2 Scenario

Urban performance currently depends not only on a city's endowment of hard infrastructure, but also, and increasingly so, on the availability and quality of communication infrastructures and services. ICT has a strategic role in this scenario, since it can optimize the distribution of knowledge and offer advanced technologies to create innovative services. The SAaaS is poised to be a very strategic solution to provide infrastructure for Smart City applications, where the involvement of both SN and mobiles would be a peculiar advantage of our vision, as in urban settings each kind of source can at a minimum serve for Cloudbursting or, in certain contexts, even augment the other due to diversification of resources and behaviors, i.e. motion patterns, user preferences, metadata, etc. In particular mobiles bring issues of random engagement to the Cloud that need to be addressed by means of volunteer mechanisms, and lead naturally to a multi-tenant scenario with regards to resource ownership. Moreover smart techniques for node management (e.g. energy optimization) are required when it comes to mobiles. Sensing resources are needed to satisfy application requirements that can continuously change to the environment conditions and the events occurring. This can be translated into a demand of high adaptability to the infrastructure, which has to provide adequate resources to promptly satisfy the requests. The SAaaS infrastructure can meet such demand by enrolling and dynamically managing the resources to satisfy fluctuating requirements. Indeed, a prominent distinction between other current sensing Clouds efforts and our vision is the offering of IaaS-like services based on sensors, i.e. proper sensing infrastructure. Here abstracting means offering homogeneous treatment of, and access to, devices, e.g. exporting SensorML-enabled representation of resources on (typically) standards-alien platforms with regards to sensing, e.g. Android, iOS, TinyOS, Contiki, etc. Leveraging abstraction, virtualization brings enablement of sensing resource morphing, composition or splitting, exposing virtualized instances of sensors, that have to be provided to the customer as devices, not exposing just data, i.e. the customer is going to be able to turn virtual knobs on the sensor handles she gets access to. The feasibility of this approach lies in the customizability of the software that drives sensing resources, i.e. our ability to offer a way to inject custom-made modules (even customer-provided possibly) into the systems in order to repurpose devices, thus both an SDK and a provider-assisted field deployment system are to be provided as part of this framework. Given the tools and infrastructure, a provider has everything in place to offer contextualization services, as per IaaS definition. Last but not least, a remarkable advantage of exposing devices is the ability to offer actuating resources as well, leading to fully feedback-enabled flows, whereas pure data-driven approaches obviously fail to engage this kind of devices.

Several case studies can be developed on a SAaaS infrastructure, where resources are gathered from diverse contributing nodes (SN, smart devices, laptops, PDA, etc.), geographically scattered across sites, as shown in Fig. 1. Building upon infrastructure services offered by an SAaaS-powered provider, a wide set of applications belonging to a Smart City scenario can be developed and deployed, among which we decided to focus on smart mobility and smart surveillance systems, also involving mobile

crowdsensing as discussed in Section 4. Through these use cases we intend to demonstrate how SAaaS can meet outstanding requirements, which are enabling in terms of Smart City applications, such as devices abstraction, due to the great variety of devices; virtualization, due to customization and security needs; scalability, due to variable and not predictable expansions of Sensor Networks; flexibility, due to heterogeneous parameters to track; fault-tolerance, due to ensure services continuity.

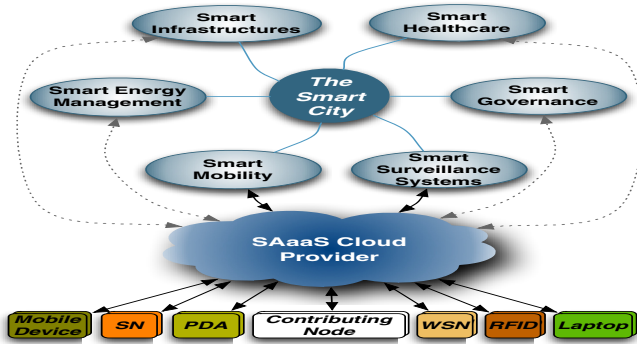


Fig. 1. Smart Cities scenario and use cases in SAaaS

This can be translated into a demand of high adaptability of the infrastructure, which has to provide adequate resources to promptly satisfy the requests. All those requirements find a natural outlet in a Cloud-based implementation that allows to adequately meet such demand by enrolling and dynamically managing the resources to satisfy fluctuating requirements. On a higher level, such an IaaS-like offering enables Software as a Service providers and end-users alike to detect potential dangerous conditions leading to emergencies (overloading conditions, black out, etc.), or to monitor crowded areas (traffic jams, sport events, safety and emergency services, etc.), just to name a few uses, where potential customers could therefore include law enforcement officials and emergency management agents.

To adopt the SAaaS approach in the Smart City scenario we can broadly identify two main phases:

1. *SAaaS Provisioning System & Infrastructure Setup* - the SAaaS provisioning system and the infrastructure should be set up in order to provide the required (virtual) sensing and actuation resources to the application, as described in Section 3.
2. *Application Deployment* - the software has to be deployed by first selecting and then customizing the (virtual) resources provided by the SAaaS infrastructure, which can be also merged to the ones owned by the application provider in a Cloudbursting hybrid-Cloud fashion, as described in Section 4.

It is important to remark that the SAaaS provider and the application/software (SaaS) provider generally don't overlap since, according to the Cloud provisioning model, infrastructure and software providers identify two different stakeholders. However,

they can also coincide, i.e. the application/software (SaaS) provider can decide to build up an SAaaS infrastructure to deploy his/her application in a private Cloud fashion. On the other hand, in order to deploy an application into a SAaaS infrastructure it is not mandatory to first build up the infrastructure, therefore the SAaaS provider & infrastructure setup phase would not be a prerequisite in that case.

3 SAaaS Provisioning System and Infrastructure Setup

This section describes the first step of the workflow described in Section 2. With this aim, an overview of the SAaaS system is provided in Section 3.1, while setup details are discussed in Section 3.2.

3.1 Overview of SAaaS Sensing Cloud

SAaaS is a paradigm aimed at developing a sensing infrastructure based on sensors and actuators from both mobiles and SNs, in order to provide virtual sensing and actuation resources in a Cloud-like fashion.

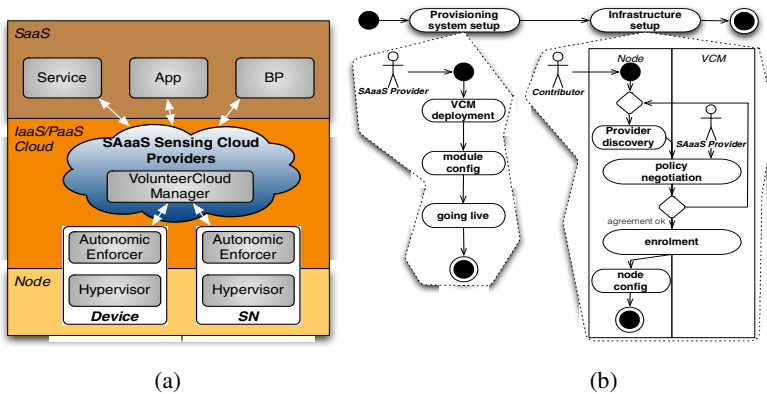


Fig. 2. Architectural schema and modules (a) and SAaaS Provisioning System and Infrastructure setup (b)

In order to build a Cloud of sensors in [1] we introduced the whole SAaaS stack and a rough schema of the architectural modules based on the functionalities to be provided and thus specifying the three main components shown in Fig. 2(a), i.e. Hypervisor, Autonomic Enforcer and VolunteerCloud Manager. The SAaaS stack and modules span to all the layers thus identified: from the highest Cloud one, providing PaaS support to SaaS software applications and services, to the IaaS infrastructure.

The lowest block of the stack, the Hypervisor one, operates at the level of a single node, where it abstracts the available sensors. The node can be a standalone resource-rich device, such as a smartphone, or it can be an embedded system which belongs to a network (such as a WSN). The main duties of the Hypervisor are: relaying commands and data retrieval, communications and networking, abstraction of devices and

capabilities, virtualization of abstracted resources, customization, isolation, intra-node management facilities, semantic labeling and thing-enabled services.

The Cloud modules, under the guise of an Autonomic Enforcer and a Volunteer-Cloud Manager, deal with issues related to the interaction among nodes, belonging to a single Cloud, when generating a Cloud of Sensors. The former is tasked with enforcement of Cloud policies, local vs. global (i.e. relayed) policy tie-breaking, subscription management, cooperation on overlay instantiation, where possible through autonomic approaches. The latter is instead in charge of exposing the generated Cloud it hides by means of Web Service interfaces, framing reward mechanisms and policies in synergy with SLA matching, to be mediated by QoS metrics and monitoring, as well as indexing duties to allow for efficient discovery of resources.

3.2 Setup

In order to build up an SAaaS system it is first necessary to configure the SAaaS provisioning system managing the requests and therefore to enroll the underlying sensing physical nodes at the basis of the SAaaS provisioning model, after abstraction and virtualization. Thus, as shown in Fig 2(b), this process can be split into two main sequential steps: provisioning system and infrastructure setup.

The former step aims at establishing the SAaaS provisioning system, instantiating, deploying and configuring the upper level components of the architecture above described, i.e. the ones belonging to the Volunteer Cloud Manager (VCM) module, up to the SAaaS provider. More specifically, as shown on the left side of Fig. 2(b), the VCM modules have to be deployed into one or more physical machines (**VCM deployment**), exposed as a Web service. The VCM deployment could be performed either in a centralized or (hopefully) in a distributed way to avoid a single point of failure.

After deployment it is necessary to configure the VCM modules (**module config**). In particular the management policies (business models, SLA policies, etc.) should be specified as well as the user interface (frontend, protocols, knowledge base, etc.). Once the system is configured and ready to the delivery the SAaaS provider has to be exposed and indexed (**going live**) in order to be discovered and selected by both interested users and potential contributors.

The second phase, the infrastructure setup, then mainly aims at enrolling sensing resources and thus it is triggered by contributors. Indeed, a potential contributor that intends to share his/her sensing resources with the SAaaS community, has to first select a specific SAaaS provider (**provider discovery**) and therefore to start the negotiation process with the latter (**policy negotiation**), negotiating on his/her involvement, the SLA/credit/reward policies and similar issues. Such process (discovery + negotiation) is reiterated upon agreement by the involved parties. In case of successful agreement the node is enrolled into the SAaaS infrastructure (**enrolment**) and thus configured as required by the SAaaS provider and agreed by the parties (**node config**).

In this way, when a number of contributing resources provide their availability to the SAaaS provider, the system is ready to accept and manage requests of (virtual) sensing and actuation resources.

4 Smart City Application Deployment in SAaaS

4.1 SAaaS Application Setup

As depicted in Fig. 3, we have five (macro) actions involved in the deployment of custom applications powered by SAaaS for any service provider to start offering her software (SaaS) or platform (PaaS) over the Web. The deployment is kickstarted with a look-up (**resource search**) for sensing resources according to predefined constraints, as per requirements of the application to be offered as a service, and thus it is triggered by the app / platform provider. This phase is followed, upon selection of the resources to be employed, by instantiation of either physical or virtualized resources (**virtual instantiation**), to be carried out Cloud-side. As soon as the app provider can get hold of handles for these instances, the setup of the whole app can commence, first by setting up (app) server components and configuring them to reference the aforementioned handles (**server-side app setup**), and then sending requests for customization of selected nodes, where needed, to be acted upon by the SAaaS machinery (**node customization**). Upon completion and acknowledgement of customization duties, it's time for the app provider to expose its web service, possibly including any frontend or UI (**web service deployment**), to be put online for customers to peruse.

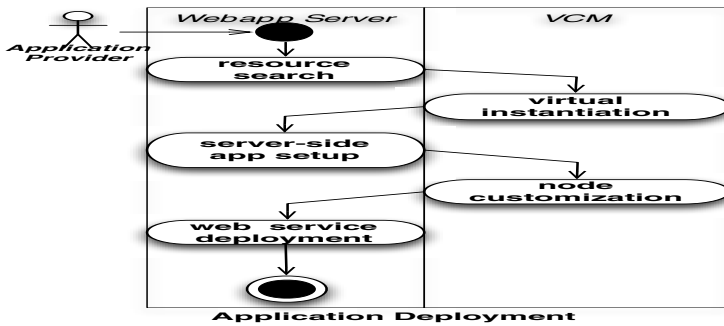


Fig. 3. Custom application deployment

4.2 Smart City UC Deployment on SAaaS

Here two compelling use cases, concerning mobility and homeland security, for SAaaS in a Smart City environment are discussed, where the novelty of our device-driven approach lends those feasibility and effectiveness. According to the proposed approach, in order to implement a Smart City application it is necessary to start from the underlying sensing infrastructure. As discussed in Section 3.2, the SAaaS

infrastructure usually is up and ready to provide virtual sensors and actuators. But, in order to show the whole process, here we assume the SAaaS provisioning system and infrastructure has not yet been established and therefore we have to first set it up. We also assume the SAaaS infrastructure gets readied to cover all the applications described in the use cases below. Following the workflow of Fig. 2(b), the stakeholder that wants to build up the SAaaS infrastructure (a company, a community, etc., e.g. the ACME company) has to first setup the SAaaS provisioning system. The ACME SAaaS provider has to thus identify the physical computing infrastructure on which to deploy and configure the SAaaS modules, including the web service frontend. Then, the enrolment phase is triggered, involving the parties interested in contributing their sensing resources, negotiating with the ACME provider their involvement as shown on the right side of Fig. 2(b). We are particularly interested in GPS, accelerometers, cameras and similar sensors to implement the use cases. Thus, once the ACME SAaaS infrastructure is ready to serve handles on virtual sensors and actuators, the use cases can be deployed on it. According to the diagram depicting the deployment of custom applications, the activities that mandate an infrastructure-oriented paradigm like SAaaS are the instantiation of virtualized sensing resources and the customization of nodes, whereas none of the data-driven approaches to sensing Clouds are able to offer this kinds of functionality.

Table 1. Smart surveillance infrastructure: key elements

<i>Kind of sensing node</i>	<i>Associated services</i>
IP cameras / CCTV	<ul style="list-style-type: none"> - Motion detection - Presence detection (e.g. heat mapping) - Environment characterization (e.g. ambient light sensing)
Environmental & Mobile	<ul style="list-style-type: none"> - Motion detection (e.g. feature extraction from captures) - Presence detection (e.g. feature extraction from captures) - Environment characterization (e.g. ambient light sensing)

Smart Surveillance Systems. In a smart city setting, an outstanding selling point of a SAaaS-enabled infrastructure can be any application that fulfills homeland security requirements. A feasible one could be based on surveillance systems where, e.g., IP cameras, in their default configuration, are employed as CCTV, yet can be repurposed for, e.g., motion or presence detection (i.e. by means of heat mapping) as well as environmental characterization (i.e. aided by ambient light sensing and such), once enabled under the guise of Infrastructure as a Service, as soon as ACME has the infrastructure (e.g. IP cameras) enrolled and set up. This kind of flexibility can be achieved by means of customization of the virtualized instances of the devices being exposed, i.e., “beaming” software modules to the embedded platforms hosting the environment these cameras would be running upon, therefore shifting many higher-level duties to the smart sensing block. The customer will have the choice to move the knob in either way to set one’s own favorite tradeoffs in terms of, e.g., higher bandwidth consumption coupled with lower processing on the embedded system vs. lower communication overhead while carrying on more (remote-side) computational duties, and so on. Once

again, none of the above could be possible in a data-driven environment, as a prerequisite for repurposing resources is access to node customization facilities. The nodes, and their associated services, to be deployed for such use case are specified in Table 1.

Smart Traffic Control. One more offering leveraging SaaS in a Smart City context could be in terms of traffic control facilities, by means of a Mobile Crowd Sensing (MCS) approach. It's quite obvious that mobile handsets can help a lot in a traffic control applications, as those devices are usually equipped with lots of relevant sensors, like those detecting position (GPS, WiFi, GSM) and motion (gyros), and are usually enabled for (always-on) bidirectional communication with the Internet, and with neighboring handsets as well.

Table 2 summarizes the parameters set of this use case.

Table 2. MCS-powered TC parameters set

<i>Observed / Controlled Parameters</i>	<i>Source (sensing)</i>	<i>Target (actuation)</i>
Geoposition (latitude, longitude, altitude)	mobile	-
Relative position (distance plus direction, orientation)	mobile	-
Motion detection (direction, orientation, rotation)	mobile	-
Scene capture	mobile	-
Human-Computer Interaction interfaces	-	mobile
Mobile platform notification system	-	mobile (virtual)

As information for traffic behavior profiling would come from users on the road, their mobility is key. Dealing with this kind of infrastructure i.e., privately owned mobiles, requires at the very least volunteer engagement and node management techniques, as pointed out before, but our efforts go far beyond, as we want to offer sensing Infrastructure under a Cloud provisioning scheme, leveraging what is a natural outcome of the scenario, i.e., multi-tenancy of resources, to be offered by one or more providers, tasked with gathering, managing and offering device-like resources to customers. Any effective road monitoring scheme would need at least a way to search (i.e., by area), filter and select sensors, including repurposed ones (i.e., cameras, with matching software to take shots autonomously, e.g., when certain events occur) by virtue of a Cloud-like provisioning model. At last, no full featured traffic control system may afford to lack mechanisms to influence traffic patterns, in our example to be implemented by feeding relevant info back to the contributors (e.g., road users) by means of (virtual) actuators, e.g., on-board signaling devices of any kind (buzzers, haptic interfaces) or other (platform) notification system piggybacking. Not to mention the possibilities disclosed by virtualizing those same devices, e.g., composing gyros and compass to expose a gyrocompass, or morph a camera into a motion detector, thus packaging up the virtual instances of sensing/actuation devices for the task at hand. Virtualization is thus achieved by means of contextualization, as unlocked by (virtual) device driver reconfiguration capabilities, where deployment takes the shape of provider-assisted packaging and delivery of customized modules for field use. In this case our ACME provider has to set up a Cloud made up of lots of mobiles for this

kind of use case to be feasible. Here there are many facilities unique to the SAaaS approach being leveraged: whether we talk about virtualization, exploited on many levels, about actuating resources or even about customization, there's a lot of requirements for enabling such a scenario, and no way to tackle any of these issues by means of a data-driven approach, where they are totally out of scope.

5 Conclusions

Combining sensing resources and Cloud infrastructure is appealing if challenging, especially when tackling the problem from a novel IaaS-like perspective. SAaaS aims at bridging the gap between sensing pervasive scenarios and service-oriented provisioning models through a device-driven approach offering facilities such as virtualization and customization of sensing and actuation devices.

In this work we have focused on a Smart City scenario, comprising two use cases, i.e. Surveillance Systems and Traffic Control, to put SAaaS unique features up on display, and validate the usefulness of our approach against quite a compelling and trendy technology domain, the former all the more relevant when compared to other solutions for sensing Clouds, where much needed facilities are out of scope and outright unfeasible.

We believe that, as ongoing development efforts progress in parallel with those related to applications for any of these use cases, we will have the chance to further investigate and obtain proofs about our claims.

References

1. Distefano, S., Merlino, G., Puliafito, A.: Sensing and actuation as a service: A new development for clouds. In: Proceedings of the 2012 IEEE 11th International Symposium on Network Computing and Applications, NCA 2012, pp. 272–275. IEEE Computer Society, Washington, DC (2012)
2. Galache, J., Santana, J., Gutierrez, V., Sanchez, L., Sotres, P., Munoz, L.: Towards experimentation-service duality within a smart city scenario. In: 2012 9th Annual Conference on Wireless On-demand Network Systems and Services (WONS), pp. 175–181 (January 2012)
3. Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., Morris, R.: Smarter cities and their innovation challenges. *Computer* 44(6), 32–39 (2011)
4. Su, K., Li, J., Fu, H.: Smart city and the applications. In: 2011 International Conference on Electronics, Communications and Control (ICECC), pp. 1028–1031 (September 2011)
5. Lee, J., Baik, S., Lee, C.: Building an integrated service management platform for ubiquitous cities. *Computer* 44(6), 56–63 (2011)
6. Li, Z., Chen, C., Wang, K.: Cloud computing for agent-based urban transportation systems. *IEEE Intelligent Systems* 26(1), 73–79 (2011)
7. Mitton, N., Papavassiliou, S., Puliafito, A., Trivedi, K.S.: Combining cloud and sensors in a smart city environment. *EURASIP J. Wireless Comm. and Networking* 2012, 247 (2012)

Author Index

- Ahmed, Nasir 402
Anvesh, T. 443
- Bai, Yuqin 141
Ben Yahia, Nesrine 61
Bevilacqua, Vitoantonio 601
Bhardwaj, Arpit 86
Bu, Fanliang 621
- Cafarchia, Domenico 601
Cai, Guixian 192
Cao, Wei 421
Cao, Yongchun 141
Cao, Zhiqiang 376
Chakrabarty, Kankana 325
Chen, Guangyi 385
Chen, Jeng-Fung 207
Chen, Peng 315
Chen, Shanshan 354
Cui, Han-Guo 612
- Dai, Gui-Ping 368
Dai, Xing 612
Deng, Suping 629
Ding, Shifei 96
Distefano, Salvatore 638
Dong, Jing 411
Du, Ji-Xiang 427
Du, Xiaojiang 285
Duc, Tran Minh 536
- Fang, Xianmei 156
Fu, Weichao 546
- Gao, Guanglai 265
Gao, Xiaobo 156
Ge, Hao 192
Goel, Garima 392
Guan, Qingxiao 411
Guo, Delong 71
Guo, Jiaxiang 331
Gupta, Phalguni 443, 453, 576, 584, 593
- Han, Jie 331
Hao, Yuheng 241
- He, Linlin 507
He, Tingting 112
He, Ying-yu 163
He, Zunliang 421
Henda, Ben Ghezala 61
Hernández, Danilo Cáceres 556
Hoang, Van-Dung 556
Hu, Min 498
Hu, Rong 8
Huang, HuaJuan 96
Huang, Min 217
Huang, Rui 376
Huang, Wenjun 55
- Islam, Saiful 593
Ivona, Fabio 601
- Jain, Shubham 453
Jiang, Lin 507
Jiang, Yanjun 227
Jo, Kang-Hyun 556
- Kang, Hee-Jun 526, 536
Kang, Soojung 259
Kaushik, Vandana Dixit 576
Khan, Hameed Ullah 402
Kobayashi, Kunikazu 21
Krishnan, Sridhar 385
Kuremoto, Takashi 21
- Li, Hongyu 546
Li, Jia-guo 482
Li, Ji-Rui 47
Li, Ke 315
Li, Kun 8
Li, Qingbao 275
Li, Shenghong 192
Li, Tong 346
Li, Yuanyuan 77
Li, Yuhua 392
Li, Zheng-Min 612
Li, Zhi-ping 39
Li, Zuo-Cheng 8
Liao, Xiangyun 331
Lin, Pengxiang 112

- Liu, Fuyong 199
 Liu, Hong 31
 Liu, Jianxun 150
 Liu, Meng 295
 Liu, Wei 498
 Liu, Wenhong 241
 Liu, Xia 285
 Liu, Yang 39, 482
 Lou, Song-Tao 47
 Lui, Shufen 490
 Luo, Liping 362
 Luo, Xiaobin 71
 Lv, Jun 346
- Ma, Jinwen 183
 Ma, Lin-Hai 106
 Ma, Yanzhui 199
 Mabui, Shingo 21
 Maguire, Liam 392
 Mandi, Soumya 584
 Marino, Francescomaria 601
 McLoone, Sean 392
 Meng, Huazhi 362
 Meng, Qingfang 354
 Merlino, Giovanni 638
 Min, Hai 516
 Ming, Zhong 629
 Mo, Yuanbin 199
 Modi, Mangat Rai 593
- Narjès, Bellamine Ben Saoud 61
 Nie, Ru 96
 Nigam, Aditya 443, 453
 Niu, Ben 629
 Niu, Xiaopeng 275
- Obayashi, Masanao 21
 Ou, Renxia 339
- Pan, Zhenghua 128, 173
 Pillai, Rajesh R. 576
 Puliafito, Antonio 638
- Qian, Bin 8
 Qu, Zhao Yang 305
- Ryu, Keun Ho 305
- Shang, Li 462
 Shao, Yabin 141
- Shi, Fanhuai 472
 Shi, Yanbin 31, 339
 Shi, Yanli 339
 Shin, Ilhoon 259
 Song, Bitna 259
 Song, Xueyan 227
 Song, Yun 421
 Sun, Jian 31
 Sun, Lan 120
 Sun, Weitao 346
 Sun, Zhan-li 462
- Tan, Tieniu 411
 Tang, Shao-fang 163
 Tao, Tongtong 192
 Tao, Wenlin 490
 Tiwari, Aruna 86
 Tiwari, Kamlesh 584
 Tsurusaki, Tetsuya 21
- Van, Mien 526
- Wan, Cong 234, 250
 Wang, Anhui 183
 Wang, Cong 250
 Wang, Cuirong 234, 250
 Wang, Fei-Zhang 612
 Wang, Fenglin 354
 Wang, Haiming 234
 Wang, Huaqing 315
 Wang, Jimao 241
 Wang, Jing 106
 Wang, Ling 305
 Wang, Wei 275
 Wang, Xiao-Feng 516
 Wang, Xiaohua 498
 Wang, Xingwei 217
 Wang, Yiyi 621
 Wang, Yong 39
 Wang, Zhibo 507
 Wei, Hongxi 265
 Weng, Xiaokang 275
 Wu, Di 546
 Wu, Guan-feng 482
 Wu, Junchao 285
 Wu, Xin 241
 Wu, Xing 427
- Xiang, Jiang 295
 Xiang, Zheng 295

- Xie, Liangjun 376
Xie, Wenfang 385
Xie, Zhihua 435
Xu, Liangfeng 498
Xu, Yong 482
Xue, Hongtao 315
- Yan, Bijin 472
Yan, Tingqin 490
Yang, Cheng 331
Yi, Yuefei 507
You, Shengyu 507
You, Zhuhong 629
Yu, Haixia 339
Yuan, Ying 234, 250
Yuan, Zhiyong 331
Yue, Jiguang 472
- Zhai, Chuan-Min 427
Zhang, Chang-Sheng 8
Zhang, Chengbo 217
Zhang, Cheng-zhi 39
Zhang, Hongli 285
Zhang, Huanhuan 141
- Zhang, Lin 546
Zhang, Li-Ping 612
Zhang, Xinfeng 546
Zhang, Xixiang 150
Zhang, Yi-Gang 516
Zhang, Yong 112
Zhang, Yu 1
Zhao, Jiexin 173
Zhao, Yongjia 385
Zhao, Zhong-Qiu 106
Zheng, Wen-Bin 47
Zhou, Bing 482
Zhou, Changxiong 490
Zhou, Tie Hua 305
Zhou, Wei 362
Zhou, Weidong 354
Zhou, Xinzhou 295
Zhou, Yongquan 71
Zhou, Zhigang 285
Zhou, Zhimin 421
Zhu, Na 250
Zhu, Yanjuan 566
Zhu, Yanqiao 183
Zhu, Zexuan 629