Fu Lee Wang
Hepu Deng
Yang Gao
Jingsheng Lei (Eds.)

Artificial Intelligence and Computational Intelligence

LNAI 6319

International Conference, AICI 2010
Sanya, China, October 2010
Proceedings, Part I

Springer

Lecture Notes in Artificial Intelligence 6319

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Fu Lee Wang Hepu Deng Yang Gao
Jingsheng Lei (Eds.)

Artificial Intelligence and Computational Intelligence

International Conference, AICI 2010
Sanya, China, October 23-24, 2010
Proceedings, Part I

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Fu Lee Wang
Caritas Francis Hsu College
Department of Business Administration
18 Chui Ling Road, Tseung Kwan O, Hong Kong, China
E-mail: pwang@cihe.edu.hk

Hepu Deng
RMIT University
School of Business Information Technology
City Campus, 124 La Trobe Street, Melbourne, Victoria 3000, Australia
E-mail: hepu.deng@rmit.edu.au

Yang Gao
Nanjing University
Department of Computer Science
Nanjing 210093, China
E-mail: gaoy@nju.edu.cn

Jingsheng Lei
Nanjing University of Posts and Telecommunications
School of Computer
Nanjing 210003, China
E-mail: leijs@njupt.edu.cn

Library of Congress Control Number: 2010936499

CR Subject Classification (1998): I.2, F.1, J.3, F.2, K.4.4, D.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-16529-X Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16529-0 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

The 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI 2010) was held October 23–24, 2010 in Sanya, China. The AICI 2010 received 1,216 submissions from 20 countries and regions. After rigorous reviews, 105 high-quality papers were selected for publication in the AICI 2010 proceedings. The acceptance rate was 8%.

The aim of AICI 2010 was to bring together researchers working in many different areas of artificial intelligence and computational intelligence to foster the exchange of new ideas and promote international collaborations. In addition to the large number of submitted papers and invited sessions, there were several internationally well-known keynote speakers.

On behalf of the Organizing Committee, we thank Hainan Province Institute of Computer and Qiongzhou University for its sponsorship and logistics support. We also thank the members of the Organizing Committee and the Program Committee for their hard work. We are very grateful to the keynote speakers, invited session organizers, session chairs, reviewers, and student helpers. Last but not least, we thank all the authors and participants for their great contributions that made this conference possible.

October 2010

Fu Lee Wang
Hepu Deng
Yang Gao
Jingsheng Lei

Organization

Organizing Committee

General Co-chairs

Qing Li	City University of Hong Kong, China
Mingrui Chen	Hainan University, China

Program Committee

Co-chairs

Hepu Deng	RMIT University, Australia
Yang Gao	Nanjing University, China

Local Arrangement

Chairs

Zhuang Li	Qiongzhou University, China

Proceedings

Co-chair

Fu Lee Wang	Caritas Francis Hsu College, China
Jingsheng Lei	Nanjing University of Posts and Telecommunications, China

Publicity

Chair

Lanzhou Wang	China Jiliang University, China

Sponsorship

Chair

Zhiyu Zhou	Zhejiang Sci-Tech University, China

Program Committee

Adi Prananto	Swinburne University of Technology, Australia
Adil Bagirov	University of Ballarat, Australia
Ahmad Abareshi	RMIT University, Australia
Alemayehu Molla	RMIT University, Australia
Andrew Stranier	University of Ballarat, Australia
Andy Song	RMIT University, Australia
An-Feng Liu	Central South University, China
Arthur Tatnall	Victoria University, Australia
Bae Hyeon	Pusan National University, South Korea
Baoding Liu	Tsinghua University, China
Carmine Sellitto	Victoria University, Australia
Caroline Chan	Deakin University, Australia
CheolPark Soon	Chonbuk National University, South Korea
Chowdhury Morshed	Deakin University, Australia
Chung-Hsing Yeh	Monash University, Australia
Chunqiao Tao	South China University, China
Costa Marly	Federal University of Amazonas, Brazil
Craig Parker	Deakin University, Australia
Daowen Qiu	Zhong Shan University, China
Dat Tran	University of Canberra, Australia
Dengsheng Zhang	Monash University, Australia
Edmonds Lau	Swinburne University of Technology, Australia
Elspeth McKay	RMIT University, Australia
Eng Chew	University of Technology Sydney, Australia
Feilong Cao	China Jiliang University, China
Ferry Jie	RMIT University, Australia
Furutani Hiroshi	University of Miyazaki, Japan
Gour Karmakar	Monash University, Australia
Guojun Lu	Monash University, Australia
Heping Pan	University of Ballarat, Australia
Hossein Zadeh	RMIT University, Australia
Ian Sadler	Victoria University, Australia
Irene Zhang	Victoria University, Australia
Jamie Mustard	Deakin University, Australia
Jeff Ang Charles	Darwin University, Australia
Jennie Carroll	RMIT University, Australia
Jenny Zhang	RMIT University, Australia
Jian Zhou T	Tsinghua University, China
Jingqiang Wang	South China University, China
Jinjun Chen	Swinburne University of Technology, Australia
Joarder Kamruzzaman	Monash University, Australia
Kaile Su	Beijing University, China
Kankana Chakrabaty	University of New England, Australia
Konrad Peszynski	RMIT University, Australia
Kuoming Lin	Kainan University, Taiwan

Lemai Nguyen	Deakin University, Australia
Leslie Young	RMIT University, Australia
Liping Ma	University of Ballarat, Australia
Luba Torline	Deakin University, Australia
Maple Carsten	University of Bedfordshire, UK
Maria Indrawan	Monash University, Australia
Peter Shackleton	Victoria University, Australia
Philip Branch	Swinburne University of Technology, Australia
Pradip Sarkar	RMIT University, Australia
Qiang Li	University of Calgary, Canada
Ravi Mayasandra	RMIT University, Australia
Richard Dazeley	University of Ballarat, Australia
Sanming Zhou	University of Melbourne, Australia
Santoso Wibowo	RMIT University, Australia
Schetinin Vitaly	University of Bedfordshire, UK
Shengxiang Yang	University of Leicester, UK
ShyhWei Teng	Monash University, Australia
Siddhi Pittayachawan	RMIT University, Australia
Stephen Burgess	Victoria University, Australia
Sungshin Kim	Pusan National University, South Korea
Syed Nasirin	Brunel University, UK
Tae-Ryong Jeon	Pusan National University, South Korea
Tayyab Maqsood R	MIT University, Australia
Tony Zhang	Qingdao Univesity, China
Vanessa Cooper	RMIT University, Australia
Wei Lai	Swinburne University of Technology, Australia
Wei Peng	RMIT University, Australia
Weijian Zhao	China Jiliang University, China
Xiaodong Li	RMIT University, Australia
Xiaohui Zhao	Swinburne University of Technology, Australia
Yan-Gang Zhao	Nagoya Institute of Technology, Japan
Yang-Cheng Lin	National Dong Hwa University, Taiwan
Yi-Hua Fan	Chung Yuan Christian University Taiwan, Taiwan
Yuan Miao	Victoria University, Australia
Yubin Zhong	Guangzhou University, China
Yubo Yuan	China Jiliang University, China
Yuefeng Li	Queensland University of Technology, Australia
Zhaohao Sun	University of Ballarat, Australia
Zhichun Wang	Tianjin University, China

Table of Contents – Part I

Applications of Artificial Intelligence

Application of RBF Neural Network in Short-Term Load Forecasting ... 1
Yongchun Liang

Efficient Large Image Browser for Embedded Systems 10
Yuanyuan Liu, Zhiwei He, Haibin Yu, and Jinbiao Liu

Formalizing Ontology-Based Hierarchical Modeling Process of Physical World 18
Nan Wang, Dantong OuYang, and Shanwu Sun

Automated Problem Solving

Satisfiability Degree Analysis for Transition System 25
Yang Zhao and Guiming Luo

Research on Optimization Design of Bottom Angle of Drag-Reducing Structure on the Riblets Surface 33
Qi-feng Zhu, Bao-wei Song, and Peng Wang

Towards Analysis of Semi-Markov Decision Processes 41
Taolue Chen and Jian Lu

Automatic Programming

Stability of Equilibrium Solution and Periodical Solution to Cohen-Grossberg Neural Networks 49
Jingsheng Lei, Ping Yan, and Teng Lv

Exponential Synchronization of Delayed Fuzzy Cohen-Grossberg Neural Networks with Reaction Diffusion Term 57
Teng Lv and Ping Yan

Magnetic Field Extrapolation Based on Improved Back Propagation Neural Network 64
Li-ting Lian, Chang-han Xiao, Sheng-dao Liu, Guo-hua Zhou, and Ming-ming Yang

Sparse Deep Belief Net for Handwritten Digits Classification 71
Jiongyun Xie, Hongtao Lu, Deng Nan, and Cai Nengbin

Data Mining and Knowledge Discovering

Multisensor Image Fusion Using a Pulse Coupled Neural Network 79
Yi Zheng and Ping Zheng

Real-Time Performance Reliability Assessment Method Based on Dynamic Probability Model 88
Cheng Hua, Qing Zhang, Guanghua Xu, and Jun Xie

Decomposing Data Mining by a Process-Oriented Execution Plan 97
Yan Zhang, Honghui Li, Alexander Wöhrer, Peter Brezany, and Gang Dai

An Efficient Distributed Subgraph Mining Algorithm in Extreme Large Graphs 107
Bin Wu and YunLong Bai

Spatio-Temporal Clustering of Road Network Data 116
Tao Cheng and Berk Anbaroglu

A Sampling Based Algorithm for Finding Association Rules from Uncertain Data 124
Qian Zhu, Donghua Pan, and Guangfei Yang

Distributed AI and Agents

Multi-agent System Collaboration Based on the Relation-Web Model ... 132
Maoguang Wang, Hong Mei, Wenpin Jiao, Junjing Jie, and Tingxun Shi

Research on a Novel Multi-agents Dynamic Cooperation Method Based on Associated Intent 145
Weijin Jiang, Xiaoling Ding, Yuhui Xu, Wang Chen, and Wei Chen

Expert and Decision Support Systems

Automobile Exhaust Gas Detection Based on Fuzzy Temperature Compensation System 153
Zhiyong Wang, Hao Ding, Fufei Hao, Zhaoxia Wang, Zhen Sun, and Shujin Li

Technical Efficiency Analysis of New Energy Listed Companies Based on DEA 161
Chong Gao, Jian-ze Zhang, and Xiao-dong Li

Fuzzy Logic and Soft Computing

Research on Differential Evolution Algorithm Based on Interaction Balance Method for Large-Scale Industrial Processes of Fuzzy Model ... 169
Dakuo He, Yuanyuan Zhao, Lifeng Wang, and Hongrui Chang

Fuzzy Control System Design for Solar Energy with PCM Tank to Fresh Air Conditioning ... 177
Jun Yang, Ailin Xiao, and Lili Wang

Intelligent Information Fusion

An Efficient Method for Target Extraction of Infrared Images ... 185
Ying Ling and Xingjin Mao

Second Order Central Difference Filtering Algorithm for SINS/GPS Integrated Navigation in Guided Munitions ... 193
Lei Cai and Xuexia Zhang

Adaptive Signal Processing for ARX System Disturbed by Complex Noise ... 201
Yulai Zhang and Guiming Luo

Abstraction for Model Checking the Probabilistic Temporal Logic of Knowledge ... 209
Conghua Zhou, Bo Sun, and Zhifeng Liu

Intelligent Scheduling

Research on Double-Objective Optimal Scheduling Algorithm for Dual Resource Constrained Job Shop ... 222
Li Jingyao, Sun Shudong, Huang Yuan, and Niu Ganggang

MDA Compatible Knowledge– Based IS Engineering Approach ... 230
Audrius Lopata and Martas Ambraziunas

Intelligent Signal Processing

ARIMA Signals Processing of Information Fusion on the Chrysanthemum ... 239
Lanzhou Wang and Qiao Li

Noise Uncertainty Study of the Low SNR Energy Detector in Cognitive Radio ... 248
Guoqing Ji and Hongbo Zhu

Machine Learning

A Realtime Human-Computer Ensemble System: Formal Representation and Experiments for Expressive Performance 256
Tetsuya Mizutani, Shigeru Igarashi, Tatsuo Suzuki, Yasuwo Ikeda, and Masayuki Shio

A New Smooth Support Vector Machine 266
Jinjin Liang and De Wu

Convergence of GCM and Its Application to Face Recognition 273
Kai Li, Xinyong Chen, Nan Yang, and Xiuchen Ye

Designing a Multi-label Kernel Machine with Two-Objective Optimization 282
Hua Xu and Jianhua Xu

Collision Detection Algorithm in Virtual Environment of Robot Workcell 292
Qian Ren, Dongmei Wu, Shuguo Wang, Yili Fu, and Hegao Cai

Machine Vision

A First Step towards Hybrid Visual Servoing Control Based on Image Moments 301
Xiaojing Shen, Dongmei Huang, and Xiaoxia Qin

A Novel Motion Detection Approach for Large FOV Cameras 311
Hongfei Yu, Wei Liu, Bobo Duan, Huai Yuan, and Hong Zhao

Large Scale Visual Classification via Learned Dictionaries and Sparse Representation 321
Zhenyong Fu, Hongtao Lu, Nan Deng, and Nengbin Cai

Semi-supervised Nearest Neighbor Discriminant Analysis Using Local Mean for Face Recognition 331
Caikou Chen, Pu Huang, and Jingyu Yang

Multi-agent Systems

Decentralized Cohesive Motion Control of Multi-agent Formation in 3-Dimensional Space 339
Ran Zhao, Yongguang Yu, and Guoguang Wen

A Model for Cooperative Design Based on Multi-agent System 349
Hua Chen, Jun Zhao, and Bo Sun

Natural Language Processing

Scaling Up the Accuracy of Bayesian Classifier Based on Frequent Itemsets by M-estimate 357
Jing Duan, Zhengkui Lin, Weiguo Yi, and Mingyu Lu

Aggressive Dimensionality Reduction with Reinforcement Local Feature Selection for Text Categorization 365
Wenbin Zheng and Yuntao Qian

Neural Networks

3D Shape Representation Using Gaussian Curvature Co-occurrence Matrix 373
Kehua Guo

Nonlinear System Identification of Bicycle Robot Based on Adaptive Neural Fuzzy Inference System 381
Xiuli Yu, Shimin Wei, and Lei Guo

Transmission: A New Feature for Computer Vision Based Smoke Detection 389
Chengjiang Long, Jianhui Zhao, Shizhong Han, Lu Xiong, Zhiyong Yuan, Jing Huang, and Weiwei Gao

A Novel Features Design Method for Cat Head Detection 397
Hua Bo

Time Serial Model of Rock Burst Based on Evolutionary Neural Network 406
Wei Gao

Multilayer Perceptron Network with Modified Sigmoid Activation Functions 414
Tobias Ebert, Oliver Bänfer, and Oliver Nelles

Pattern Recognition

Kernel Oblique Subspace Projection Approach for Target Detection in Hyperspectral Imagery 422
Liaoying Zhao, Yinhe Shen, and Xiaorun Li

Text-Independent Voice Conversion Based on Kernel Eigenvoice 432
Yanping Li, Linghua Zhang, and Hui Ding

An Effective Method for SAR Automatic Target Recognition 440
Ying Li and Hongli Gong

Semi-supervised Learning by Spectral Mapping with Label Information 448
Zhong-Qiu Zhao, Jun Gao, and Xindong Wu

Drop Fingerprint Recognition Based on Self-Organizing Feature Map 455
Jie Li, Qing Song, Yuan Luo, and Cunwei Zou

Nonlinear Complementarity Problem and Solution Methods 461
Longquan Yong

Robotics

A New Path Planning Method for a Shape-Shifting Robot 470
Mengxin Li, Ying Zhang, TongLin Liu, and Chengdong Wu

Pedestrian Gait Classification Based on Hidden Markov Models 479
Weihua Wang and Zhijing Liu

L-Infinity Norm Minimization in the Multiview Triangulation 488
Yang Min

Erratum

An Efficient Method for Target Extraction of Infrared Images E1
Ying Ling and Xingjin Mao

Author Index 495

Table of Contents – Part II

Applications of Computational Intelligence

A New Fault Detection Method of Induction Motor 1
Chuanbo Wen and Yun Liang

A Method to Identify Damage of Roof Truss under Static Load Using Genetic Algorithm 9
Ying Wang, Jianxin Liu, Fengying Shi, and Jun Xiao

Non-linear Improvement on Hydraulic Pump and Motor Models Based on Parameter Optimization Algorithms 16
Anlin Wang, Binnan Yue, Kaifei Jiang, and Xiaotian Li

Diurnal and Seasonal Changes in Stem Water Content of Single Yulan Magmolia Tree 24
Hailan Wang and Yandong Zhao

Reliability Analysis on Wing Structures under the Gust Load 31
Xiaozhou Ning, Yunju Yan, Kangkang Qu, and Zhilao Li

Prediction Interval on Spacecraft Telemetry Data Based on Modified Block Bootstrap Method 38
Jiahui Luan, Jian Tang, and Chen Lu

Application of Sleep Scheduling Mechanism in Three-Dimensional Environment 45
Tongneng He and Peijun Chen

Dimensions of E-commerce Benefits as Perceived by Businesses 52
Xibao Zhang

Nonlinear Analysis of a Hybrid Optimal Velocity Model with Relative Velocity for Traffic Flow 58
Tao Liu and Lei Jia

Biomedical Informatics and Computation

Insertion Force of Acupuncture for a Computer Training System 64
Ren Kanehira, Weiping Yang, Hirohisa Narita, and Hideo Fujimoto

Classifying Motor Imagery EEG Signals by Iterative Channel Elimination according to Compound Weight 71
Lin He, Zhenghui Gu, Yuanqing Li, and Zhuliang Yu

Automatic Reference Selection for Quantitative EEG Component Interpretation: Cross Spectrum Analysis Based on Bipolar EEG 79
Bei Wang, Xingyu Wang, Akio Ikeda, Takashi Nagamine, Hiroshi Shibasaki, Takenao Sugi, and Masatoshi Nakamura

Mixed Numerical Integral Algorithm for Deformation Simulation of Soft Tissues 87
Hui Liang and MingYong Shi

Multiple Sequence Alignment Based on ABC_SA 98
Xiaojun Xu and Xiujuan Lei

TDMA Grouping Based RFID Network Planning Using Hybrid Differential Evolution Algorithm 106
Xiang Gao and Ying Gao

An Improved PSO-SVM Approach for Multi-faults Diagnosis of Satellite Reaction Wheel 114
Di Hu, Yunfeng Dong, and Ali Sarosh

Research of Long-Term Runoff Forecast Based on Support Vector Machine Method 124
Yong Peng and Zhi-chun Xue

Fuzzy Computation

Application of Latin Hypercube Sampling in the Immune Genetic Algorithm for Solving the Maximum Clique Problem 134
Benda Zhou and Minghua Chen

The Selection of Sales Managers in Enterprises by Fuzzy Multi-criteria Decision-Making 142
Yu-Jie Wang, Chao-Shun Kao, and Li-Jen Liu

Towards the Impact of the Random Sequence on Genetic Algorithms ... 152
Yongtao Yu and Xiangzhong Xu

A New Pairwise Comparison Based Method of Ranking LR-fuzzy Numbers 160
Mingxin Zhang and Fusheng Yu

A Fuzzy Assessment Model for Traffic Safety in City: A Case Study in China 168
Zhizhong Zhao, Boxian Fu, and Ning Zhang

Genetic Algorithms

An Optimization Model of Site Batch Plant Layout for Infrastructure Project 175
Kwan-Chew Ng, Jing Li, Chen-Xi Shi, and Qian Li

Damping Search Algorithm for Multi-objective Optimization Problems 185
Jia Ji, Jinhua Peng, and Xinchao Zhao

Pruned Genetic Algorithm 193
Seyyed Mahdi Hedjazi and Samane Sadat Marjani

Immune Computation

A New Computational Algorithm for Solving Periodic Sevendiagonal Linear Systems 201
Xiao-Lin Lin and Ji-Teng Jia

Local Weighted LS-SVM Online Modeling and the Application in Continuous Processes 209
Lijuan Li, Hui Yu, Jun Liu, and Shi Zhang

Information Security

A Cellular Automata Based Crowd Behavior Model 218
Dalong Wang, Ngai Ming Kwok, Xiuping Jia, and Feng Li

A Novel Watermark Technique for Relational Databases 226
Hazem El-Bakry and Mohamed Hamada

Intelligent Agents and Systems

A Cell-Phone Based Brain-Computer Interface for Communication in Daily Life 233
Yu-Te Wang, Yijun Wang, and Tzyy-Ping Jung

DCISL: Dynamic Control Integration Script Language 241
Qingshan Li, Lei Wang, Hua Chu, and Shaojie Mao

Mapping Multi-view Architecture Products to Multi-agent Software Architecture Style 249
Zhongxue Li, Haiming Zhong, and Xike Wang

Nature Computation

ID-Based Authenticated Multi-group Keys Agreement Scheme for Computing Grid 259
Xiaofeng Wang and Shangping Wang

Dynamic Path Planning of Mobile Robots Based on ABC Algorithm . . . 267
Qianzhi Ma and Xiujuan Lei

Urban Arterial Traffic Coordination Control System 275
Jianyu Zhao, Diankui Tang, Xin Geng, and Lei Jia

A Semiparametric Regression Ensemble Model for Rainfall Forecasting Based on RBF Neural Network 284
Jiansheng Wu

Particle Swarm Optimization

A Modified Particle Swarm Optimizer with a Novel Operator 293
Ran Cheng and Min Yao

An AntiCentroid-oriented Particle Swarm Algorithm for Numerical Optimization 302
Xinchao Zhao and Wenbin Wang

Comparison of Four Decomposition Algorithms for Multidisciplinary Design Optimization 310
Peng Wang, Bao-wei Song, and Qi-feng Zhu

Multilevel Image Thresholding Selection Using the Artificial Bee Colony Algorithm 318
Ming-Huwi Horng and Ting-Wei Jiang

Automatic Rule Tuning of a Fuzzy Logic Controller Using Particle Swarm Optimisation 326
Gu Fang, Ngai Ming Kwok, and Dalong Wang

An Efficient Differential Evolution Algorithm with Approximate Fitness Functions Using Neural Networks 334
Yi-shou Wang, Yan-jun Shi, Ben-xian Yue, and Hong-fei Teng

Probabilistic Reasoning

Evaluate the Quality of Foundational Software Platform by Bayesian Network 342
Yuqing Lan, Yanfang Liu, and Mingxia Kuang

Triangle Fuzzy Number Intuitionistic Fuzzy Aggregation Operators and Their Application to Group Decision Making 350
Dongfeng Chen, Lei Zhang, and Jingshan Jiao

Statistical Analysis of Wireless Fading Channels 358
Hao Zhang, Yong Liu, and Junxiang Gao

Discretization Method of Continuous Attributes Based on Decision Attributes 367
Yingjuan Sun, Zengqiang Ren, Tong Zhou, Yandong Zhai, and Dongbing Pu

Empirical Research of Price Discovery for Gold Futures Based on Compound Model Combing Wavelet Frame with Support Vector Regression 374
Wensheng Dai, Chi-Jie Lu, and Tingjen Chang

Author Index 383

Application of RBF Neural Network in Short-Term Load Forecasting

Yongchun Liang

Department of Electrical and Information
Hebei University of Science and Technology
Shijiazhuang, China
lycocean@163.com

Abstract. A Radius Basic Function (RBF) neural network is proposed for the power load forecasting. RBF neural network can meet nonlinear recognition and process predition of the dynamic system, and has better adaptability to dynamic forecasting and prediction problem in mechnism. The RBF centres are determined by the orthogonal least squared (OLS) learning procedure. The effectiveness of the model and algorithm with the example of power load forecasting have been proved and approximation capability and learning speed of RBF neural network is better than BP neural network.

Keywords: RBF network; Load forecasting; Power system.

1 Introduction

Load forecasting is foundation of power system planning and operation. Accurate load forecasting provides reasonabelly guide to arange the start- stop of generation unit. The process of load forecasting is to analyze the history load data and pre-estimate the development of load with known power system and meteorological information. It is a tradtional research problem in power system [1,2].

Load forecasting can be divided into long-term forecasting, medium-term forecasting and short-term forecasting. Long-term forecasting is used to provide the basic data of source and network and determine the operation mode and annual maitenance plan. Medium-term forecasting is used to determine the month maitenance plan. Short-term forecasting is used to plan the power generation and anlyze real-time security and dispatch the power. This paper focus on the short-term forecasting method.

In recent years, artificial intelligence methods have been used widely in load forecasting, such as expert system, artificial neural network, and so on. Artificial neural network is considerated to be one of the most future forecasting method. Artificial neural networks (ANN) are systems that are inspired by biological neural networks. The characteristic feature of an ANN is that it considers the accumulated knowledge acquired through training examples and responds to new events in the most appropriate manner, on the basis of experiences gained through training. This means that ANN have the ability to learn the desired input-output mapping based on training examples, without looking for an exact mathematical model. Once an

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 1–9, 2010.

appropriate ANN is properly trained, it will contain a representation of the nonlinearity of the desired mapping between the inputs and outputs. The ability of ANNs to learn complex nonlinear input/output relationships have motivated researchers to apply ANNs for solving model-free nonlinear problems related to various fields, including those of power-system engineering.

Currently, the multilayer feedforward network with a back-propagation (BP) training algorithm is the most widely used ANN model for various applications, and has also been used for load forecasting. However, despite its wide applications, BP network has a number of deficiencies, such as slow training and local minimum. In order to overcomes these defects and improve the diagnosis accuracy, genetic algorithm (GA) is used in artificial neural network. GA-BP network can overcome the defect of entrapment in local optimum and search the optimum in the whole scope. But the training rate is still slow and auto-adaptability is not very good [3,4].

The arbitrariness in the design of a neural network trained by a BP algorithm, for its number of hidden layers and the number of neurons per hidden layer, is removed in the RBF neural network by having only one hidden layer and the number of neurons in the hidden layer optimised for a particular system set of input pattern of data. RBF neural network can approximate any continuos function with arbitrary accuracy. Approximation capability and learning speed of RBF neural network is better than BP neural network.

2 RBF Neural Network

RBF neural network is proposed by J.Moody and C.Darken. RBF neural network has a feed forward structure consisting of three layers, such as an input layer, a nonlinear hidden layer and a linear output layer, as shown in Fig.1. The hidden nodes are the radial basis function units and the output nodes are simple summations. The number of input , output and hidden nodes are ni, nh, no, respectively. This particular architecture of RBF neural network has proved to directly improve training and performance of the network. Any of the function, namely spline, multiquadratic and Gaussian function, may be used as a transfer function for the hidden neurons. The Gaussian RBF, which is the most widely used one, has been considered for the proposed load forecasting [5-6].

The response of the jth hidden neuron to the input can be expressed as:

$$\phi_j(x, x_j) = \exp(-\frac{1}{\sigma_j^2}\left\|x-\mu_j\right\|^2) \quad 1 \le j \le n_h \tag{1}$$

where μ_j is the centre for the jth hidden neuron , σ_j is the spread of the Gaussian function and $\|\ \|$ denotes the Euclidian norm. the output of each node in the output layer is defined by

$$f_i(X) = \sum_{j=1}^{n_h} \phi_j(\left\|X-\mu_j\right\|)\theta_{ji} \quad 1 \le i \le n_o \tag{2}$$

where X is the input vector and θ_{ji} represents the weight from the jth hidden node to the ith output node.

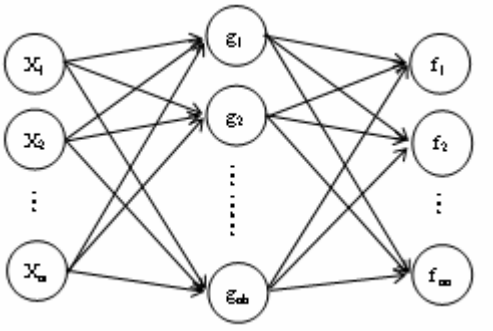


Fig. 1. Structure of RBF neural network

The performance of an RBF neural network depends on the choice of the values of the centers. In the proposed work, the simulation studies have been carried out by means of MATLAB's neural network toolbox, which makes use of the orthogonal least squares (OLS) learning procedure for determining the RBF centers.

The OLS procedure can be implemented by considering a RBF ϕ with a spread σ and introducing an error term ε in (2), which can be rewritten as

$$d_i(X) = \sum_{j=1}^{n_h} \phi_j(\|X - \mu_j\|)\theta_{ji} + \varepsilon_i \quad 1 \le i \le n_o \tag{3}$$

where d_i is the desired output of the *i*th output node, and then maximising the error reduction ratio by orthogonal LS principle.

The task of network learning is to choose appropriate centers and determine the corresponding weights based on a given set of training inputs and outputs. The RBF network centers are selected from training data $\{X(1), \cdots\cdots, X(N)\}$ and it is a problem of subset model selection.

The full model is defined by considering all the training data as candidates for centers. In the matrix form (3) can be expressed as

$$D = \phi\Theta + E \tag{4}$$

where

$$D = [d_1, \cdots\cdots, d_{no}] \tag{5}$$

$$d_i = [d_i(1) \cdots\cdots d_i(N)] \quad i = 1, \cdots\cdots, n_o \tag{6}$$

$$\phi = [\phi_1 \cdots\cdots \phi_M] \tag{7}$$

$$\phi_i = [\phi_i(1) \cdots\cdots \phi_i(N)]^T \quad i = 1, \cdots\cdots, N \tag{8}$$

$$\Theta = [\theta_1 \cdots\cdots \theta_{no}] \tag{9}$$

$$\theta_j = [\theta_j(1) \cdots\cdots \theta_j(N)]^T \quad j = 1 \cdots\cdots M \tag{10}$$

$$E = [\varepsilon_1 \cdots\cdots \varepsilon_{no}] \tag{11}$$

$$\varepsilon_i = [\varepsilon_i(1) \cdots\cdots \varepsilon_i(N)]^T \quad i = 1 \cdots\cdots n_o \tag{12}$$

The parameter matrix Θ can be solved by using the LS principle. The regressor ϕ_i form a set of basis vectors. These bases are generally correlated. An orthogonal transformation can be performed to transfer from the set of ϕ_i into a set of orthogonal basis vectors. This can be achieved by decomposing ϕ into

$$\phi = WA \tag{13}$$

where

$$A = \begin{bmatrix} 1 & & & \\ & \cdot & \beta_{pq} & \\ & & \cdot & \\ & 0 & & \cdot \\ & & & 1 \end{bmatrix} \begin{matrix} p = 1, \cdots, M-1 \\ q = 1, \cdots, M \end{matrix} \tag{14}$$

and

$$W = [w_1 \cdots w_M] \tag{15}$$

with the orthogonal columns that satisfy

$$w_i^T w_j = 0 \tag{16}$$

where $i \neq j$

$$w_i^T w_i = h_i^2 \tag{17}$$

Equation (4) can be rewritten as

$$D = WA\Theta + E \tag{18}$$

With OLS solution

$$A\Theta = G = \begin{bmatrix} g_{11} & \cdot & g_{1no} \\ g_{M1} & \cdot & g_{Mno} \end{bmatrix} \tag{19}$$

The standard Gram-Schmidt method may be used to derive A and G. Thus, Θ can be solved from (19).

In general, the number of all candidate regressors N can be very large, and an adequate modelling may only require n_h regressors, where n_h may be smaller than N. The significant regressors can be selected using the OLS procedure. As the error matrix E is orthogonal to W, the error reduction ratio due to w_m can be found by

$$[err]_m = \left(\sum_{i=1}^{n_o} g_{mi}^2 \right) w_m{}^T w_m \,/\, trace\left(D^T D\right) \quad 1 \le m \le N \tag{20}$$

Based on this ratio, the significant regressors can be selected in a forward regression procedure. At the mth step, a candidate regressor is selected as the mth regressor of the subset network, if it produces the largest value of $[err]_m$ from among the rest of the N-m+1 candidates. The selection is terminated when

$$1 - \sum_{m=1}^{n_h} [err]_m < \rho \tag{21}$$

where ρ is a chosen small tolerance.

In the context of a neural network, the OLS learning procedure choose the radial basis function centres $\{\mu_1, \mu_2, \cdots, \mu_N\}$ as a subset of training data vectors $\{X(1), \cdots\cdots, X(N)\}$, where $n_h < N$. The centres are determined one by one in a well-defined manner, following the Gram-Schimidt orthogonalisation procedure until a network of adquate performance is constructed. In case the output of an individual jth hidden neuron ϕ_j^N is consistently much less than $\max\{\phi_j^N\}$, for j=1 to N, then the jth hidden neuron may be ignored or removed.

3 Load Forecasting Using RBF Neural Network

In this paper, 12 sets of load data are measured every 2 hours in one day. Otherwise, power load is depend on environmental factors. The highest temperature, the lowest temperature and the weather characteristic are measured too. When the weather information is known, the load data of forecast day can be calculated by RBF neural network.

In proposed RBF neural network, there are 15 input nodes, 25 hidden nodes and 12 output nodes. All 12 sets of load data and 3 environmental factors are the input vector of RBF neural network. The 12 sets of load data of next day are the output vector of RBF neural network. The load datas and weather information of 11 days are shown in Table 1. The power load and weather features from 2008.7.1 to 2008.7.10 are the input training data and the power load from 2008.7.2 to 2008.7.11 are the output training data. All the datas have been normalized.

Table 1. Power load and weather feathures

Date	Power Load			Weather Features
2008-7-1	0.2452	0.1466	0.1314	
	0.2243	0.5523	0.6642	
	0.7015	0.6981	0.6821	
	0.6945	0.7549	0.8215	
2008-7-2	0.2217	0.1581	0.1408	0.2415 0.3027 0
	0.2304	0.5134	0.5312	
	0.6819	0.7125	0.7265	
	0.6847	0.7826	0.8325	
2008-7-3	0.2525	0.1627	0.1507	0.2385 0.3125 0
	0.2406	0.5502	0.5636	
	0.7051	0.7352	0.7459	
	0.7015	0.8064	0.8156	
2008-7-4	0.2106	0.1105	0.1243	0.2216 0.2701 1
	0.1978	0.5021	0.5232	
	0.6819	0.6952	0.7015	
	0.6825	0.7825	0.7895	
2008-7-5	0.2115	0.1201	0.1312	0.2352 0.2506 0.5
	0.2019	0.5532	0.5736	
	0.7029	0.7032	0.7189	
	0.7019	0.7965	0.8025	
2008-7-6	0.2335	0.1322	0.1534	0.2542 0.3125 0
	0.2214	0.5623	0.5827	
	0.7198	0.7276	0.7359	
	0.7506	0.8092	0.8221	
2008-7-7	0.2368	0.1432	0.1653	0.2601 0.3198 0
	0.2205	0.5823	0.5971	
	0.7136	0.7129	0.7263	
	0.7153	0.8091	0.8217	
2008-7-8	0.2342	0.1368	0.1602	0.2579 0.3099 0
	0.2131	0.5726	0.5822	
	0.7101	0.7098	0.7127	
	0.7121	0.7995	0.8126	
2008-7-9	0.2113	0.1212	0.1305	0.2301 0.2867 0.5
	0.1819	0.4952	0.5312	
	0.6886	0.6898	0.6999	
	0.7323	0.7721	0.7956	
2008-7-10	0.2005	0.1121	0.1207	0.2234 0.2799 1
	0.1605	0.4556	0.5022	
	0.6553	0.6673	0.6798	
	0.7023	0.7521	0.7756	
2008-7-11	0.2123	0.1257	0.1343	0.2314 0.2977 0
	0.2079	0.5579	0.5716	
	0.7059	0.7145	0.7205	
	0.7401	0.8019	0.8136	

The training process is shown in Fig.2.

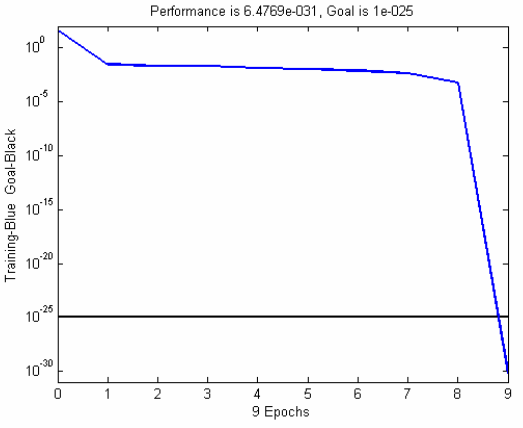


Fig. 2. Convergence Curve of RBF network

The weather features of 2008.7.12 are [0.2317 0.2936 0]. The power loads of 2008.7.12 calculated by RBF neural network are [0.2352 0.1401 0.1581 0.2249 0.5689 0.5893 0.7168 0.7237 0.7379 0.7363 0.8112 0.8259].

4 Analysis

The squared difference between the output data of BP network and measured load data is shown in Fig.3.

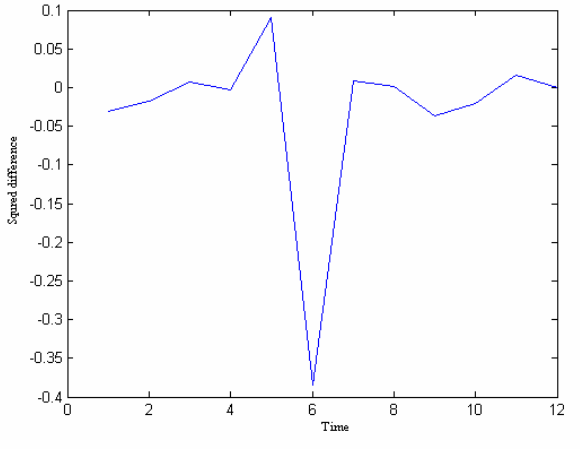


Fig. 3. Error Curve of BP network

The squared difference between the output data of RBF network and measured load data is shown in Fig.4.

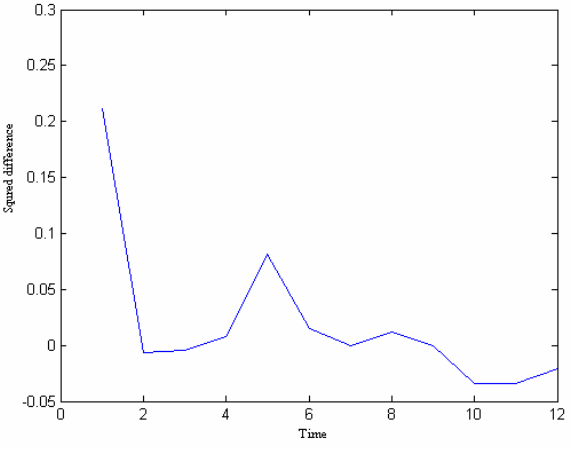


Fig. 4. Error Curve of RBF neural network

The maximum squared difference of BP network is about 3.8. the maximum square difference of RBF network is aoubt 2.2. this proved that RBF network is better than BP network for load forecasting.

5 Conclusion

By comparing with the BP neural network algorithm, we conclude that our proposed model is better and more efficient than the traditional BP network methods. The RBF network provides an effective and convenient tool especially for the load forecasting. The forecasting accuracy is high.

Acknowledgments. This paper is supported by Natural Science Foundation of Hebei Province (Grand No: E2009000719).

References

1. Chongqing, K., Qing, X., Boming, Z.: Review of Power System Load Forecasting and Its Development. Automation of Electric Power System 28, 1–11 (2004)
2. Songhuai, D.: Load Forecasting Method and Its Influencing Factors. Rural Electrification 4, 17–18 (2001)
3. Jie, Y.: The Application of Neural Network to Electric Load Forecasting. North-East Power Technique 12, 47–48 (2007)

4. Yaonian, L., Deyou, Y., Songling, P.: Power System Short-term Load Forecasting Based on Empirical Mode Decomposition and Dynamic Neural Network. Advanced Technology of Electrical Engineering and Energy 27, 13–17 (2008)
5. Wang, L.: RBF Neural Network Predictive Control for Coagulant Dosage. In: Proceedings of the 5th World Congress on Intelligent Control and Automation, Hangzhou, P.R.China, pp. 2608–2611 (June 2004)
6. Mahanty, R.N., Dutta Gupta, P.B.: Application of RBF Neural Network to Fault Classification and Location in Transmission Lines. IEE Proc.-Gener. Transactions on Distribution 151, 201–212 (2004)

Efficient Large Image Browser for Embedded Systems*

Yuanyuan Liu, Zhiwei He, Haibin Yu, and Jinbiao Liu

Department of electronic information
Hangzhou Dianzi University
No.2 Street, Xiasha Higher education Zone
Hangzhou, China
{liuyuanyuan,zwhe,shoreyhb,ab}@hdu.edu.cn

Abstract. Images obtained by digital cameras nowadays become larger and larger in size. The quality of the obtained images is very excellent, but on the other hand, they are hard to be displayed on such embedded systems as mobile phones, digital photo frames, etc., which have both very limited memory size and very small screen size. In this paper, an efficient large image browser for embedded systems is described. A set of approaches based on pixel resampling technology are proposed to make large images to be displayed effectively. The proposed browser consumes much less memory than many famous image browser softwares. Experimental evaluations of the proposed mechanism indicate that it is efficient and effective for large image display on embedded systems.

Keywords: large image; browser; embedded systems.

1 Introduction

Statistics show that, among all the received information of a human being, more than 70 percent of them are obtained through the sense of sight. So images and videos are two kinds of very important media in people's daily life, more and more people express, share and exchange information through images and videos. According to peoples' increasing requirement and with the incremental development of CMOS sensor and VLSI technology, the pixel resolution of digital cameras and digital vidicons become more and more high, e.g., a digital camera with a 10Mega-pixels CMOS sensor is very common nowadays. The quality of the captured images is accordingly very high, but the memory consumption of such images is also very high, especially when they are browsed by an image viewer. Suppose we have a full color image captured by a 10-mega-pixels digital camera, whose resolution is 3648*2736, then it consumes a memory size of about 30M bytes for the image to be fully displayed. There is of course no problem for today's personal computer to display such an image, because it can have a memory of more than 1G bytes, with many

* This work was supported by Foundation of Department of Education of Zhejiang Province with Grant No. Y200907573 and partially supported by Zhejiang Province Nature Science Foundation of China with Grant No. Y1090881.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 10–17, 2010.

softwares run on it that can be applied , such as Adobe Photoshop and ACDSee etc. But if the hardware platform changes from PC to embedded systems such as the widely used mobile phones, digital photo frames, handheld PCs, where image browser is also a very important application, problems may occur because the size of the internal memory is very limited in such systems, and the screen is much smaller. For example, the typical embedded system—an operating system supported mobile phone— may have a internal memory size of 32M bytes, a LCD touch screen of the size of 4.3 inches with a resolution of 800*480 pixels. As the operating system run on the mobile phone also needs to consume memory, the memory left for other applications is much more less than 30M bytes. So measures should be taken before the large images can be browsed in such memory limited embedded systems.

Many efforts have been put on image adaptation and related fields from quite different aspects, which cover from the coding process to image semantic analysis.

For instance, the ROI coding scheme and Spatial/SNR scalability in JPEG 2000 [1] have provided a functionality of progressive encoding and display. It is useful for fast database access as well as for delivering different resolutions to terminals with different capabilities.

Jin Li [2] etc. implemented an interactive large image browser base on the JPEG 2000 standard. During the browsing process, the user specifies a region of interest (ROI) with certain spatial and resolution constraint, the browser only downloads the portion of the compressed bit stream that covers the current ROI, and the download is performed in a progressive fashion so that a coarse view of the ROI can be rendered very quickly and then gradually refined as more and more bit stream arrives. In the case of the switch of ROI, e.g., zooming in/out or panning around, the browser uses existing compressed bit stream in cache to quickly render a coarse view of the new ROI, and in the same time, request a new set of compressed bit stream corresponding to the updated view.

The JPEG 2000 based methods requires the images to be encoded according to the JPEG 2000 standard first, but for the images obtained by most of today's digital cameras, the stored image format is just '.jpg' which are encoded according to the JPEG standard. Because the conversion of the image format from JPEG to JPEG 2000 can only be done with a JPEG decoding followed by a JPEG 2000 encoding, there is no way to do the conversion directly. With the same memory limitation problem, apparently it is impossible to convert the image format from JPEG to JPEG 2000 on the embedded system before the images can be browsed. And it is also not convenient to require the users to convert the image format on a PC before it can be browsed on the embedded system. It can be concluded that the JPEG 2000 based methods is not suitable for the browse of large images in an embedded system.

Xing Xie etc. proposed an attention model based image adaptation approach in [3], [4]. Instead of treating an image as a whole, they manipulated each region-of-interest in the image separately, which allowed delivery of the most important region to the client when the screen size is small. In [5], they further extended the image attention model [3] to include a time constraint to calculate an optimal browsing path to maximize the information throughput under limited space and time. The main problem of the attention model based methods is that they consumed an average of more than 8 seconds to browse an image, which is actually a bit longer.

In this paper, a resampling based method is proposed and used to browse large images in embedded systems. The organization of the paper is as follows. Section 1 is an introduction of the whole work, Section 2 gives the framework of the whole system, Section 3 shows the detail description of the proposed methods, Section 4 gives the test results and the conclusion is given in Section 5.

2 The System Framework

For most of today's digital cameras, the captured images are usually stored in the format of "JPG" or "RAW". Those images with a postfix of "JPG" in the file name are encoded according to the JPEG standard, so they should be firstly decoded to obtain the color information at each pixel position before they can be browsed; while the images with a postfix of "RAW" in the file name are not compressed at all, the color information at each pixel can be read directly from the file. Thanks to the official reference implementation of the JPEG decoder, we need not to decode the whole image at one time to obtain the color information of pixels, but each time we can only decode one line to obtain the color information of pixels in that line. This is really very good news for us, otherwise we can not browse the "JPEG" encoded images at all on memory limited embedded systems for the decoding procedure consumes too much memory. With this in hand, we can just treat both of these two types of images with no differences.

For an input image, a complete system framework of the proposed approach to browse it in an embedded system is shown in Fig. 1.

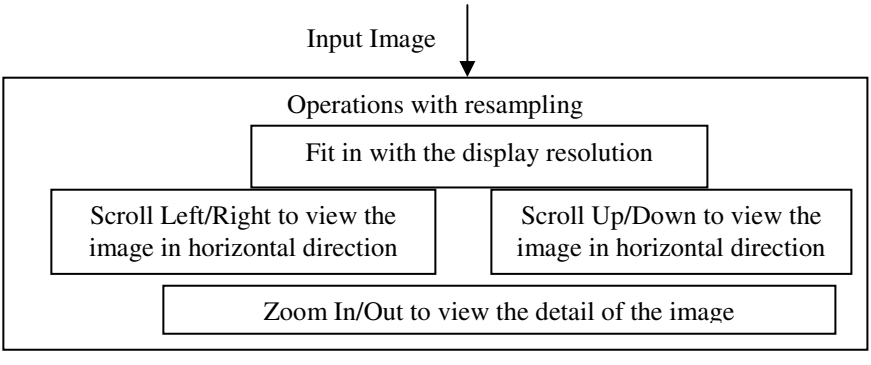


Fig. 1. System framework of the resampling based image browsing

The whole framework contains four types of operations: Fit in with the display size, Scroll Left/Right, Scroll Up/Down and Zoom In/Out. Among all the operations, "Fit in with the display size" is a basic one and also a built-in one; it is needed by all other three operations and called automatically whenever needed. "Scroll Left/Right" is used to display the content of the image which is in a left/right position compared to the current position. Similarly, "Scroll Up/Down" is used to display the content of the image which is in an up/down position compared to the current position. The "Zoom In/Out" operation can let the user view the image in a finer/coarser manner.

All the operations, except for the "Fit in" one, are user-interactive. Each time an operation is made, a resamping procedure to the original image is needed.

3 Operations Detail

Suppose the display resolution is $W_s \times H_s$, which means that it can display W_s pixels in the horizontal direction and H_s pixels in the vertical direction. On the other hand, we have an image which has a resolution of $W \times H$ to be browsed. We also suppose that both W and H are much larger than W_s and H_s. Under this condition, the detail of the operations for image browsing is as follows. One thing important which should be mentioned is that all the operations mentioned before are disabled when no image is opened.

3.1 The Fit in Operation

The "Fit-in" operation is automatically activated when a new image is browsed. The process of fitting in with the display resolution is depicted in Fig.2.

The image is opened in a binary mode; the lines are read in one by one and only those lines which is an integer times of the value W / W_s are preserved for further handling. The preserved lines are then resampled every H / H_s pixels. Finally, an image with the resolution of $W_s \times H_s$ is obtained and can be displayed on the screen.

When the image is browsed in this manner, the "Zoom-In" operation is activated for the user to view the details of the image. Other operations are still disabled at this time.

3.2 The Zoom In/Out Operation

In our proposed methodology, the "Zoom-In/Out" operation has very little differences with the "Fit-in" operation except that one more parameter z_r, which stands for "Zoom Ratio", controls the resampling procedure. The value of z_r is in the range of W / W_s to 1. Each time a "Zoom-In" operation is made, z_r is decreased to one half of its original value; on the other hand, when a "Zoom-Out" operation is made, z_r is increased to two times of its original value. If the "Zoom- Ratio" equals to 1, the "Zoom In" operation is disabled; and when we have a z_r of W / W_s, it is just the "Fit-in" case, the "Zoom Out" operation is disabled. Different from the "Fit-In" operation where we extract every W / W_s lines and pixels, we extract every z_r lines from the original images and every z_r pixels from each of the extracted line from the beginning in the "Zoom-In/Out" operation, the extraction process stops when

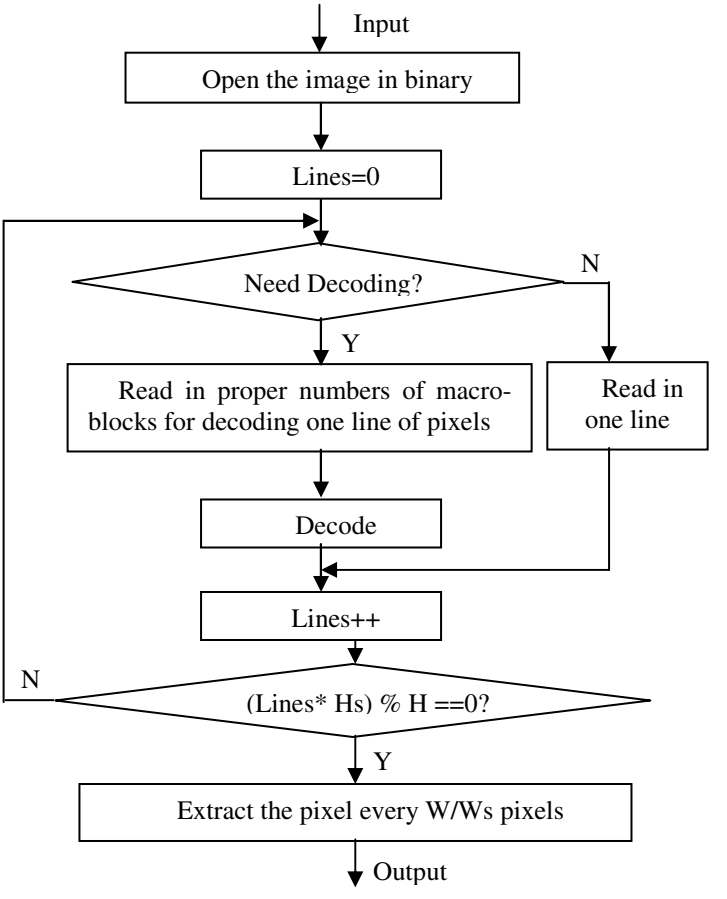


Fig. 2. The Fit-in Process when the image is firstly opened

totally H_s lines are extracted with each line containing W_s pixels. When z_r is smaller than W / W_s, the "Scroll Left/Right" and "Scroll Up/Down" operation are activated.

3.3 The Scroll Left/Right Operation

As we discussed in the preceding subsection, when the "Zoom Ratio" z_r is smaller than W / W_s, there are W / z_r pixels, which is larger than W_s, that can be extracted if we extract the lines every z_r pixels, but the screen can only display W_s pixels in

the horizontal direction, so the scroll left/right operation is activated for the user to view those part that are not displayed. The scroll position, which is defined to be s_r, is in the range between 0 and $\frac{W}{z_r} - W_s$. With this s_r in hand, the "Scroll Left/Right" operation is almost the same with the "Zoom In/Out" operation except that the resampling starts from the horizontal position of s_r other than 0, which means the pixels at the position of

$$s_r, s_r + 1 \cdot z_r, s_r + 2 \cdot z_r, ..., s_r + (W_s - 1) \cdot z_r$$

are extracted at this time.

3.4 The Scroll Up/Down Operation

The Scroll UP/Down operation is very similar to the Scroll Left/Right operation. We also define a variable s_d, which is in the range between 0 and $\frac{H}{z_r} - H_s$, to represent the scroll position in the vertical direction; and only the lines at the position of

$$s_d, s_d + 1 \cdot z_r, s_d + 2 \cdot z_r, ..., s_d + (H_s - 1) \cdot z_r$$

are extracted.

3.5 Zoom In/Out after Scroll

Suppose somebody wants to view the detail of the image at the center, he can zoom in the image as large as possible so that he can view the detail and then drag the scroll bar to the position he want, but he can also zoom in the image only to some extent and then drag the scroll bar to the position and then zoom in again if he wants. So a zoom in/out operation after scroll is also possible, and this time we need not only to change the "zoom ratio" z_r but also the scroll position s_r and s_d. Every time a zoom in operation is made, s_r and s_d should be decreased to one half of their original values, on the other hand, when a zoom out operation is made, they should be increased to two times of their original values.

4 Experimental Results

The proposed operations have been implemented with the software of Embedded Visual C++ 4.0 under the Operating System of Windows CE 5.0. The Graphical User Interface is shown in Fig. 3.5.

The image opened in Fig.3 has a resolution of 23622*11811 pixels, with each pixel occupies 1 bit. The memory consumed by the software is 9982K Bytes. It can be displayed in less than 1 second in an embedded system with a CPU run at 400 MHz and with a memory size of 64MB. The image should be re-read every time a new

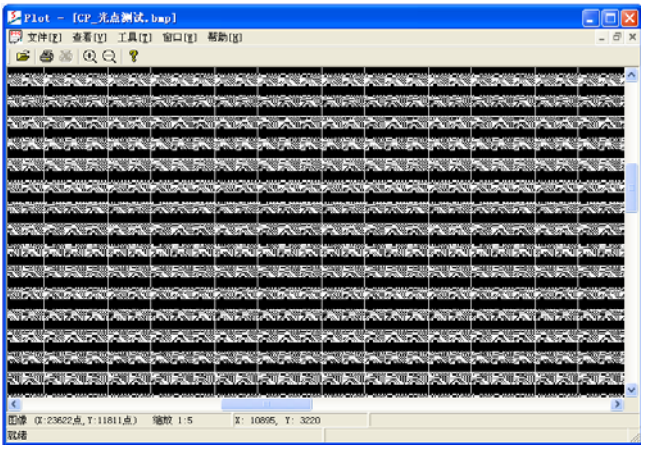


Fig. 3. GUI of the implemented software

operation mentioned above is made, but the re-display time is still less than 1 second. With a comparison, we also display the image in other popular image processing softwares such as ACDSee and Adobe Photoshop. The results are given in Table 1.

Table 1. Comparison to other popular softwares

Software	System	Memory Occupied	Display time
ACDSee 5.0	Windows XP, Pentium D 3.4GHz, 1GB Memory	16MB	3.0 s
PhotoShop 8.0	Windows XP, Pentium D 3.4GHz, 1GB Memory	120MB	<1.0s
The proposed	WinCE 5.0, PXA 270 400MHz, 64MB memory	10 MB	<1.0s

5 Conclusions

In this paper, an efficient and effective method for displaying very large images is proposed. The images are viewed in an interactive manner with several operations such as Zoom In/Out, Scroll Left/Right and Scroll Up/Down supplied. For all the operations, one basic procedure, the resampling, is needed. The starting positions in both the horizontal and the vertical directions, companioned with the zoom ratio are used for the decision how the resampling procedure should be made. Experimental results show that our framework is very suitable for applications on embedded systems.

The drawback of the proposed method is that aliasing may exist when the down-sampling ratio is big, e.g., when the image is firstly opened and fit-in with the screen.

The user can improve this by zooming in the image a bit. Another solution to the aliasing problem is that we can perform some smoothing operations to the extracted lines and pixels. For instance, we can extract several lines at one time and make a linear combination of them to obtain one smoother line, which can decrease the influence of the aliasing effect. But this will increase the memory consumption along with the display time. A tradeoff should be made then.

References

1. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 still image coding system: an overview. IEEE Trans. Consumer Electron. 46(4), 1103–1127 (2000)
2. Li, J., Sun, H.-H.: On Interactive Browsing of Large Images. IEEE Trans. on Multimedia 5(4), 581–590 (2003)
3. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. ACM Multimedia Syst. J. 9(4), 353–364 (2003)
4. Fan, X., Xie, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: Visual attention based image browsing on mobile devices. In: Proc. ICME 2003, Baltimore, MD, vol. I, pp. 53–56 (July 2003)
5. Xie, X., Liu, H., Maand, W.-Y., Zhang, H.-J.: Browsing Large Pictures Under Limited Display Sizes. IEEE Trans. on multimedia 8(4), 707–715 (2006)

Formalizing Ontology-Based Hierarchical Modeling Process of Physical World

Nan Wang[1,2], Dantong OuYang[1,*], and Shanwu Sun[3]

[1] Department of Computer Science and Technology, Jilin University, Changchun 130012
[2] Department of Information, Jilin University of Finance and Economics, Changchun, 130117
[3] Network and Laboratorial Center, Jilin University of Finance and Economics, Changchun, 130117
ouyangdantong@163.com

Abstract. The important step of model-based reasoning is to construct the abstraction model of physical world. In this paper, we introduce the concept of hierarchical ontology classes within the G-KRA model framework to realize the modeling knowledge sharing and reuse. Three kinds of ontology operators are defined and the formal process of them working is partly given. We also point out that constructing hierarchical ontology classes can be independent of the modeling process. Moreover we show how to automatically build up a hierarchical model of a world W based on the hierarchical ontology classes which is as well updated during the modeling. We expect that this work will rich the way to construct hierarchical model of physical world automatically and efficiently.

Keywords: G-KRA Model Hierarchical Ontology Classes Hierarchical Modeling Knowledge Sharing and Reuse.

1 Introduction

In resent years, researchers have paid a lot of attention to the abstraction of the physical world. The KRA model [1,2,3] is proposed to help both the conceptualization phase of a problem and the automatic application of abstraction operators. The extended model, called G-KRA model (General KRA model) [4] can represent the world from different abstraction granularities. Ontology has been proposed as a specification mechanism to enhance knowledge sharing and reuse across different applications [5] and many researchers has defined ontology in different ways [6,7,8,9,10].

In the second section of this paper, we explore the levels included in the KRA model, particularly the perception level, the language level and the theory level, so that the representing framework is simplified by knowledge sharing and reuse. We introduce the concept of hierarchical ontology class. The abstraction operators acting on the ontology classes as well as the mappings between them are also defined. In the following section, we discuss constructing process of hierarchical ontology classes and formalizing the ontology-based hierarchical modeling process of the physical world .We will make a short conclusion to present work and next-step research in the last section.

* Corresponding author.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 18–24, 2010.

2 Introducing Ontology in the KRA Model

The KRA model represents the world from four levels: Perception, Structure, Language and Theory (see [1] for details). In the G-KRA model, the abstraction model can be automatically constructed by using the abstraction objects database and the abstraction mapping defined manually. We introduce the concept of ontology and extend the abstraction objects database to Ontology Class (OC) which can be defined as OC=(E,L,T), such that E={OBJ,ATT,FUNC,REL}. The elements of E and the rest of OC mean the same as the ones in [3]. OC exists like database to realize knowledge sharing and reuse. There are two kinds of ontology class according to the constructing process: Fundamental Ontology Class (FOC), Abstraction Ontology Class (AOC). The former is generated based-on the physical world W, whereas the latter through operating on another different ontology class.

The ontology-based KRA model framework can be represented as follows: $R=(P^*,S,L^*,T^*)$, in which the three parts of P^*,L^*,T^* represent not the actual contents but the mapping to some OC, whereas OBS and S remain for being involved with the real world.

We define three kinds of ontology abstraction operators (sets): fundamental operator Ψ, entity operator set $\Omega=\{\omega_E,\omega_L,\omega_T\}$ and connection operator set $\Phi=\{\lambda_E,\lambda_L,\lambda_T\}$, to create or extend the ontology class. The operation of all the operators can be independent of modeling process. Operator Ψ applies to some physical world W and FOC, which specifies or updates FOC, that is $\Psi(W,FOC_1)=FOC_2$. The arguments of the operator set Ω are an existing ontology class and in particular, Ω combines a grouping of indistinguishable entity types of an ontology class OC_1 into one entity type in an ontology class OC_2, $\Omega(OC_1)=OC_2$. The connection operator set also acts on an ontology class and it aggregates a set of entity types of OC_1 and the relations between them to form a new compound entity type of OC_2, $\Phi(OC_1)=OC_2$.

```
---------------------------------------------------------------------------------------------------
Procedure FundamentalOperatorΨ WORLD W, ONTOLOGY-CLASS FOC
For every entity e of W
  For every e* of FOC{
  if SMP(e, e*) and BMP(e, e*) exit;// SMP and BMP are two procedures to do the
structure matching and behavior matching , see details in [11];
  type=TYPE(e);//Define a new entity type
  att=ATT(e); //Define the attribute set of e
  func=FUNC(e);//Define the behavior set of e
  rel=REL(e); //Check the entities related to e and construct the relation rel;
  Add (type,att,func,rel) to FOC.E; //Add a new entity type to FOC.E
  Add Language(e) to FOC.L;
  Add Theory(e) to FOC.T;// Add a new reasoning ontology of FOC about e
}
```

Fig. 1. Constructing the Fundamental Ontology Class

```
---------------------------------------------------------------------------------------------------------
Procedure EntityOperatorSetΩ ONTOLOGY-CLASS OC1, ONTOLOGY-CLASS OC2 //OC2=Ω(OC1)
   EntityOperatorω_E(OC1.E, OC2);
   EntityOperatorω_L (OC1.L, OC2);
   EntityOperatorω_T (OC1.T, OC2);

Procedure EntityOperatorSetΦ ONTOLOGY-CLASS OC1, ONTOLOGY-CLASS OC2//OC2=Φ(OC1)
   EntityOperatorλ_E (OC1.E, OC2);
   EntityOperatorλ_L (OC1.L, OC2);
   EntityOperatorλ_T (OC1.T, OC2);

Procedure EntityOperatorω_E ONTOLOGY-CLASS_E E, ONTOLOGY-CLASS OC2
//The operation process of operator ω_E
Do while there exist(s) unchecked entity(ies) in E{
 Entities={}; // The set of indistinguishable entity types.
   For every unchecked entity E1 of E{
   Entities={E1};
      For every unchecked entity E2 of E-{E1}
        if SMP(E1,E2) and BMP(E1,E2) Entities= Entities ∪ {E2};
  }
   //Search the indistinguishable entity types of E.

   if |Entities|>1 {
      Replicate E1 to a new entity E*;
      If there is no entity type E* in OC2, add E* to OC2;
      Update RELs of OC2.E using the RELs of E and replace every entity type E′ of
OC2.E with E* if E′∈Entities;
      Update the entity mapping EM from OC2.E to E;
  }
     //If there exit more than one indistinguishable entity types, do the combining.

   if |Entities|=1{
      Update OC2.E using E;
      Update the entity mapping EM from OC2.E to E;
  }
     //If there is no other entity type indistinguishable with E1, do the updating.
 }
---------------------------------------------------------------------------------------------------------
```

Fig. 2. Partial Process of Operators sets Ω and Φ

There are two relationships between ontology classes: entity mapping(EM) and connection mapping(CM), which make one(/more) object(/s) of ontology class OC_1 relating to one(and only one) object of OC_2. The mapping relationships can be constructed during the hierarchical process of ontology classes.

3 The Hierarchical Modeling Process of Physical World

The very important step of model-based reasoning is to construct the model of the real world. And the hierarchical model can help to reduce the complexity of the reasoning. The ontology-based KRA model framework makes it possible to automatically generate such a hierarchical model of the world W.

3.1 The Hierarchical Process of Ontology Classes

Fundamental operator Ψ acquires knowledge directly from some world W and it perceives all entities in W, including their types, attributes, behaviors, and the relations between them. See Fig 1 for details.

Both entity operator set $\Omega=\{\omega_E,\omega_L,\omega_T\}$ and connection operator set $\Phi=\{\lambda_E,\lambda_L,\lambda_T\}$ act on an existing ontology class. Fig 2 shows the partial process. We only give the algorithm that the first operator ω_E of Ω, because λ_E operates just like ω_E only not searching a set of indistinguishable entity types but finding out a set of interrelated ones. The rest operators, based on the results of ω_E *and* λ_E , are just the process which deals with variables replacing and expressions aggregating of one-order logic.

We note that while using operator λ_E, the connection mapping CM from a compound entity type to a set of interrelated entities is constructed or updated. Through alternate operating of the ontology abstraction operators (sets), the hierarchical ontology classes can be built up automatically, Fig 3 shows the process.

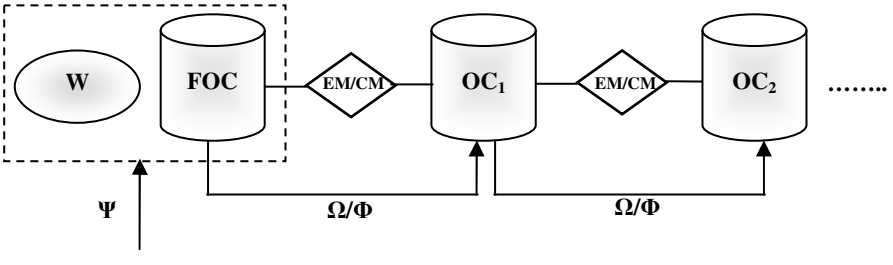


Fig. 3. Constructing the Hierarchical Ontology Classes

3.2 Hierarchical Modeling of Physical World Based-on Ontology Classes

The hierarchical ontology classes are generated or extended independently of the modeling process of the physical world. The more kinds of worlds are input, the more knowledge ontology classes can be achieved. Actually, while modeling a world W within the simplified G-KRA framework, the ontology classes can also be updated dynamically. Fig 4 shows the hierarchical modeling process of physical world based-on ontology classes.

It is noted that the definition of condition C can specify the abstraction degree of the hierarchical model. We can set a threshold value like the number of the entity types in some ontology class and if the number is lower than the threshold value, the procedure should be stopped. Fig 5 shows the hierarchical modeling process of W based-on hierarchical ontology classes.

Procedure HierarchicalModeling WORLD W, ONTOLOGY-CLASS-SET {FOC,OC_1,…,OC_n}
//Construct fundamental model R_g=(P_g*, S_g, L_g*,T_g*) by using FOC;
For every entity type e of W {
if ($\exists e_1 \in$ FOC.E and e_1 is indistinguishable with e)
Create a mapping M_e from P_g* to the entity type e_1 of FOC;
else Add the entity type e to FOC and Create a mapping M_e from P_g* to it;
}
//Create the mapping from the entity types of W to the entity types of FOC

For every type of relations r in W {
if ($\exists$ $REL_1 \in$ FOC.E and REL_1 is the same type of relation as r and no mapping on REL_1)
Create a mapping M_r in P_g* to REL_1;
else Add the relation type r to P_g* and Create a mapping M_r in P_g* to it;
}
//Create the mapping from the relation types of W to the relation types of FOC

Manually add new language and theory involved with W to FOC.L and FOC.T if need be and create mappings from L_g*, T_g* to FOC.L, FOC.T;
//Constructing the fundamental model of W

//Generate hierarchical models R_{OC1},…R_{OCn}, and what abstraction degree of the model
//is our satisfaction can be specified by a condition C defined in advance.
i=1;
do while C is satisfied {
ROC_i=ROC_{i-1}; //define ROC_0 as Rg
Adjust all mappings in R_{OCi} to OC_i according to the entity mapping EM or the connection mapping CM between R_{OCi} and R_{OCi-1};
i=i+1;
}

Fig. 4. Hierarchical Modeling Process of Physical World Based-On Ontology Classes

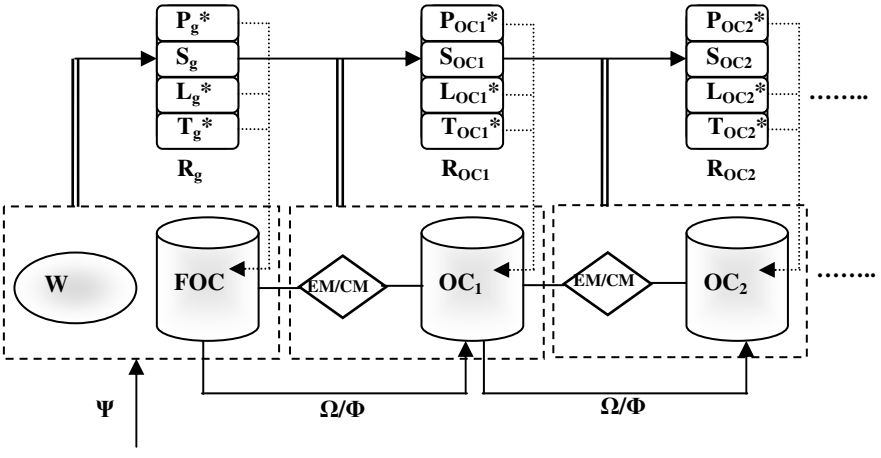


Fig. 5. Hierarchical Modeling Process of W Based-On Hierarchical Ontology Classes

4 Conclusion

This work explores how to use the concept of hierarchical ontology classes to realize the modeling knowledge sharing and reuse and then to automatically construct the hierarchical model of physical world. The KRA model framework is be simplified by replacing actual contents with the mapping to some ontology class. At the same time, the hierarchical ontology classes can be built up independently of the modeling process. Here we only concern about the static world and will continue to work on the modeling of the dynamic world in future.

Acknowledgment

The authors are grateful to the support of NSFC Major Research Program under Grant Nos. 60496320 and 60496321, Basic Theory and Core Techniques of Non Canonical Knowledge; NSFC under Grant Nos. 60973089,60773097 and 60873148; Jilin Province Science and Technology Development Plan under Grant Nos.20060532 and 20080107.Erasmus Mundus External Cooperation Window's Project (EMECW): Bridging the Gap, 155776-EM-1-2009-1-IT-ERAMUNDUS-ECW-L12.

References

1. Saitta, L., Zucker, J.: Semantic Abstraction for Concept Representation and Learning. In: Proc. SARA, pp. 103–120 (1998)
2. Saitta, L., Zucker, J.-D.: A Model of Abstraction in Visual Perception. Applied Artificial Intelligence 15(8), 761–776 (2001)
3. Saitta, L., Torasso, P., Torta, G.: Formalizing the abstraction process in model-based diagnosis. In: Tr cs, 34. Univ. of Torino, Italy (2006)

4. Shan-wu, S., Nan, W., Dan-tong, O.: General KRA Abstraction Model. Journal of Jilin University (Science Edition) 47(3), 537–542 (2009)
5. Neches, R., Fikes, R., Finin, T., Gruber, T., Senator, T., Swartout, W.: Enabling technology for knowledge sharing. AI Magazine 12, 36–56 (1991)
6. Uschold, M.: Knowledge level modelling: concepts and terminology. The Knowledge Engineering Review 13(1), 5–29 (1998)
7. Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Mars, N. (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, pp. 25–32. IOS Press, Amsterdam (1995)
8. Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43(5/6), 907–928 (1995)
9. William, S., Austin, T.: Ontologies. IEEE Intelligent Systems, 18–19 (January/February 1999)
10. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What Are Ontologies, and Why Do We Need Them?, pp. 20–25 (January/February 1999)
11. Wang, N., OuYang, D., Sun, S., Zhao, C.: Formalizing the Modeling Process of Physical Systems in MBD. In: Deng, H., Wang, L., Wang, F.L., Lei, J. (eds.) AICI 2009. LNCS, vol. 5855, pp. 685–695. Springer, Heidelberg (2009)

Satisfiability Degree Analysis for Transition System*

Yang Zhao and Guiming Luo

School of Software, Tsinghua University
Beijing 100084, China
yang-zhao08@mails.tsinghua.edu.cn, gluo@tsinghua.edu.cn

Abstract. Classical model checking is not capable of solving the situation of uncertain systems or fault-tolerant systems which occur in the real world commonly. The technique of satisfiability degree (SD) for model checking is a efficient way to solve this problem. Finite paths transition system (FPTS) for calculating satisfiability degree of a LTL logic formula is given. Then, a more general situation about discrete-time Markov chains (DTMCs) is discussed. Then a case named leader election shows the practicability of satisfiability degree for transition system, which cannot be solved by classical model checking.

Keywords: satisfiability degree, transition system, LTL.

1 Introduction

Classical logic is related certainty, i.e., any proposition is either true or false, with no other intermediate states. However, the real world does not always reflect the classical logic since there are many cases with neither absolute true nor absolute false. So many studies have used non-classical logic, such as modal logic, intuition logic, fuzzy logic, and probabilistic logic. Paper [1] analyzes the satisfiability of a proposition. The calculus of satisfiability degree is proposed, which expresses uncertainty, extends the concepts of satisfactory and contradictory propositions in propositional logic and describes how satisfactory is based on the proportion of interpretations that make the proposition true.

In this paper we want to provide a new sight of model checking to deal with the uncertain situation. In the real world, may sometimes the system can not satisfy the property totally, but it can support partially. Then classical model checking methods have difficulties to describe these cases. The conception of Satisfiability Degree is introduced into model checking, to calculate the satisfiability degree of a logic formula to express the level about a transition system satisfy the formula.

The paper is organized as follows: Section 2 introduces the definition of satisfiability degree in Propositional Logic. Section 3 discusses the satisfiability degree on transition systems, extending to the system FPTS and DTMCs. Section 4 gives a simple example to show the application of satisfiability degree on transition systems, leader election. The conclusions are given in the last section.

* This work is supported by the Funds NSFC 60973049 and 60635020.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 25–32, 2010.

2 Satisfiability Degree

Let P be a propositional formula set and Ω the global field for interpreting P, thus any formula $\varphi \in P$ and $\omega \in \Omega, \varphi(\omega) \in \{0,1\}$. If $\varphi \in P$, define a subset $\Omega\varphi \subset \Omega$ such that:

$$\Omega\varphi = \{\omega \mid \varphi(\omega) = 1, \omega \in \Omega\} \tag{1}$$

Definition 1 (Satisfiability Degree in propositional logic)
Given a propositional formula set **P** and the global interpretation field Ω, the subset is defined as above. Then, function $f: P \to [0,1]$ is called the satisfiability degree on Ω, if any formula $\varphi \in \mathbf{P}$:

$$f(\varphi) = d(\Omega\varphi)/\, d(\Omega) \tag{2}$$

where d(X) is denote as the cardinality of set X.

The most intuitive method for computing the satisfiability degree is enumeration of all the assignments using simulation as a truth table. However the complexity of this method is exponential. Three methods are presented to reduce the unnecessary costs [1-3]. One is base on the satisfiability degree properties, while another is constructed as an Extended Ordered Binary Decision Diagram (XOBDD) [2, 3]. The backtracking search algorithm for SAT is used and optimized modification also can help speed up the calculation.

3 Satisfiability Degree for Transition System

As we know, the concept of satisfiability degree on propositional logic is come up for the satisfiable properties of Propositional Logic formulas. Similarly, in field of model checking the basic idea is that whether a transition system S is a model of a logic formula F must be checked, that is, we want to know whether S can make F satisfiable.

Definition 2 (Finite paths transition system (FPTS))
Finite paths transition system is a seven-triple (S, δ, I, F, AP, L) where

- S: is a finite set of states;
- $\delta \subseteq S \to S$ is a transition relation function;
- $I \subseteq S$ is a set of initial states, which means for any s in I, there is no state t in S so that $\delta(t) = s$;
- $F \subseteq S$ is a set of final states, which means for any s in F, there exist no state t in S so that $\delta(s,) = t$;
- AP is a set of atomic propositions, and L: $S \to 2^{AP}$ is a labeling function.
- We write $\delta(s_i,) = s_j$, if s_i can transfer to s_j directly, $\delta^k(s_i) = s_j$ is written to show s_i can transfer to s_j by k steps; so the last conditions is given as follows,
- For any state s in S, there is no k exist so that $\delta^k(s) = s$, where $k \in N$;

Definition 3 (Path)
Let D be an FPTS, a path ρ in D is an sequence $(s_0, s_1, s_2 \ldots, s_n)$, since s_0 is one in I ,and s_n in F, and for any i in N, there exist $\delta(s_i) = s_{i+1}$

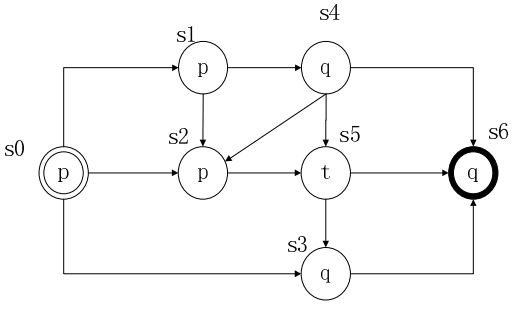


Fig. 1. A simple FPTS example

Figure 1 show an example about FPTS , AP={p, q, t}, and L is given L(s0) = L(s1) = L(s2)={p}, L(s3)=L(s4)=L(s6)={q}, L(s5)={t}. Many paths such as {s0, s2, s5, s6} and {s0, s1, s4 s5, s6} can be found, but the amount of the paths are finite, and for each path it has finite states.

On FPTS we can give the satisfiability degree definition of a LTL formula easily; the definition of linear-time temporal logic (LTL) can be found in [5]. Then for a LTL formula, we may give the definition of Satisfiability Degree on FPTS, to show the level FPTS satisfy a LTL formula.

Definition 4 (SD on FPTS)
Given an FPTS D=(S, Act, δ, I, F, AP, L), P is a LTL formula set, then the function $f_D: P \to [0,1]$ called Satisfiability Degree on D, if for any formula $\varphi \in P$:

$$f_D(\varphi) = \frac{NP(D|\varphi)}{NP(D)}, \tag{3}$$

Where NP(D) means the number of paths in D, and $NP(D|\varphi)$ is denoted as the number of paths in D which satisfy the formulaφ.

Next, the algorithm is provided to compute the SD on FPTS. we may modify the classical algorithm DFG (Deep First Search) [6] [8] to make it work. SD= (m, n) is denoted where both m and n are integers. We calculate the satisfiability degree by SD using m minus n so from the algorithm we will get SD;

Algorithm 1. Get satisfiability degree of a LTL formula on a FPTS;

```
1:  m := 0;
2:  n := 0;
3:  path :=null; // path is a stack, also represent a path
4:  for each s in S
5:    S_s= {t|δ(s,act) = t, if there exist act in Act}
6:  for each s in I
7:     {
8:     SD sd := SD-Visit(0,s);
9:     m := m + sd.m;
10:    n := n + sd.n;
11:    }
12:  return (m, n)
```

Algorithm 2. SD-Visit (s, t)

```
1:  SD sd(m, n) :=(0, 0);
2:  Push (path, t);
3:  if t is in F {
4:     sd.m := sd.m + CheckLTL(path, φ);
5:     sd.n := sd.n + 1; }
6:  else {
7:      for each u in St {
8:         SD sd2(m, n) = SD-Visit(t, u);
9:         sd.m := sd.m + sd2.m;
10:         sd.n := sd.n + sd2.n; } }
11:    Pop (path, t);
12:    return sd;
```

Algorithm 3. CheckLTL(path, φ)

```
1: if path have not satisfied φ
2:    return 0;
3: else return 1;
```

It is trivial that n is the total number of all paths in this FPTS and m represents the number of paths satisfy the formula. So from the definition, we can easily get the SD by m/n. The worst situation may cost no more than $O(2^n)$, for n represents the number of states in S. And the best situation may cost O(n), when only one path exists.

Definition 5(DTMCs)
A (labeled) discrete-time Markov chain (DTMC) is a triple D=(S, P, L) where:

- *S* is a finite set of states;
- P: $S \times S \rightarrow [0,1]$ is a stochastic matrix;
- L: $S \rightarrow 2^{AP}$ is a labeling function which assigns to each state $s \in S$ the set L(s) of atomic propositions that are valid in S

Definition 6(Paths)
Let D=(S, P, L) be a DTMC. An infinite path ρ in D is an infinite sequence $(s_0, s_1, s_2 \ldots, s_n)$ of states such that for all i>0, P(si,si+1)> 0.

A finite path σ is a finite prefix of an infinite path.

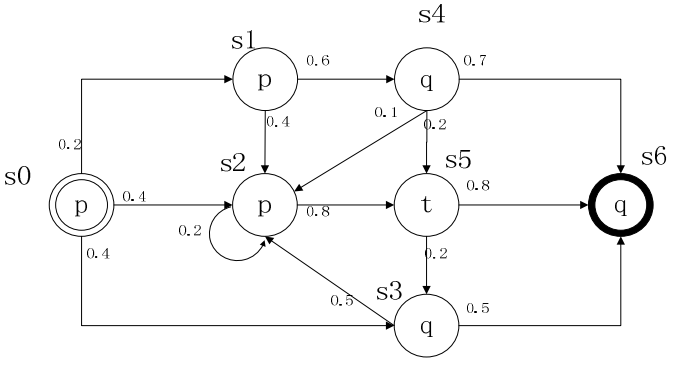


Fig. 2. A example about DTMCs

A DTMCs is given in Figure 2 which is modified from the one in figure1. The difference is that The state s2 has a self-circle and s2 can transfer to itself through s5 and s3. So obviously it can have infinite paths. And every paths has a weight.

We denote the Ω is the set of all paths in a DTMCs D, and φ is a LTL formula, let satisfied set Ωφ={ρ | ρ satisfy φ, ρ in D} a subset of Ω. And characteristic function is denoted like this:

$$X\varphi(\rho)=\begin{cases} 1 & (\text{if } \rho \in \Omega\varphi) \\ 0 & (\text{if } \rho \in \Omega \setminus \Omega\varphi) \end{cases} \tag{4}$$

Then the definition of satisfiability degree is denote as below:

Definition 7 (**SD on DTMCs**)
Let D=(S, P, L) be a DTMC, $\sum$ is a LTL formula set, then the function $f_D: \sum \to [0,1]$ called satisfiability degree on D, if for any formula $\varphi \in \sum$:

We have

$$f_D(\varphi) = \sum_{\rho \in \Omega} W(\rho) X\varphi(\rho) \tag{5}$$

We also can write $f_D(\varphi)$ as $f(\varphi|D)$.

The form of SD definition for DTMCs is very different from FPTS for the probability and infinite situation considered, so in order to compute SD we must transfer the infinite situation to finite situation (mainly caused by self-circles).

The difficulty is how to deal with self-circles. As we know LTL has some operator as X, F, G, U, W, R. For such as X, we can just check the next state to get the result, another word, and finite step must be finished, but the others we must check more states, maybe infinitely.

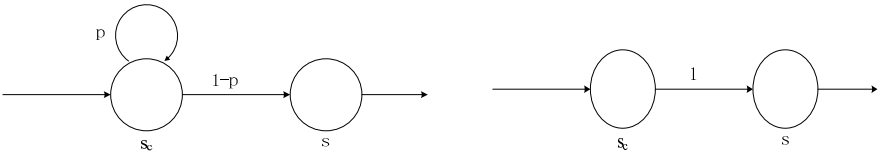


Fig. 3. Simplification of self-circles

Considering when just arriving a state s_c owning a self-circle and as it happens we just right check the formula Fφ, whereφ is a formula. If $L(s_c)$ can make φ right, then the path

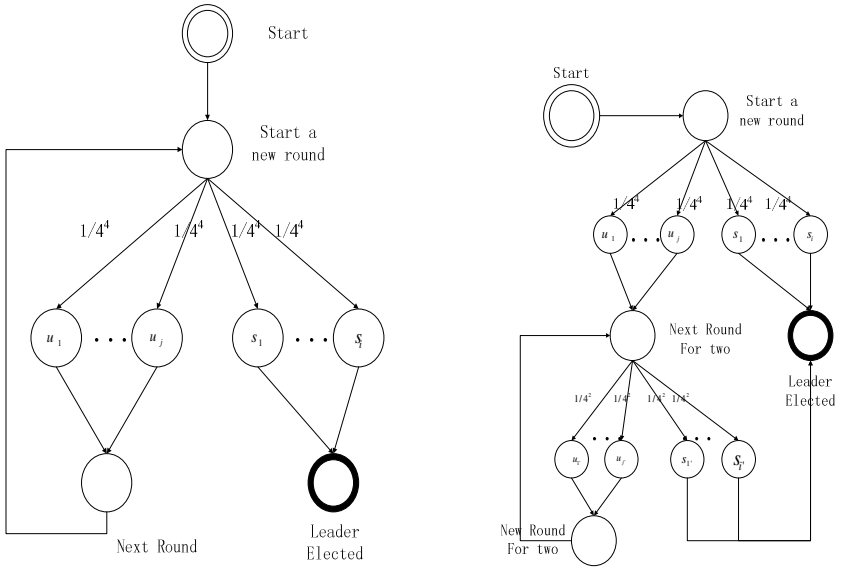
$$* s_c s *, * s_c s_c s *, \quad * \underbrace{s_c \ldots s_c}_{i} s *$$

all satisfy the formula φ where i can be any positive integer and * represents the other states before sc and after s. Vice versa. So if wrong, we can stop checking the next sc, and back-track to the previous one. And if right, we just calculate the sum $\sum_{i=0}^{\infty} p^i(1-p)=1$, and continue to check the next one. So the situation of figure 3(a) can be changed to the figure3 (b) as below.

4 Applications

In this chapter a case called Leader Election [7] is introduced to show how our satisfiability degree on transition system works. Without loss of generality, we consider four electors select one leader from their selves. The one who will be the leader must have more than the half, that is, must be selected by three people (including himself). Every person is a candidate while a elector, who has a unique id selected from {1,2,3,4} We have two ways: First, everyone can vote one (including himself) randomly, then check the result, if someone has three or more supporters, then he win, otherwise we enter the next round and do it again, until someone become the leader. Second method, we do the same thing on the first round, but if no one is selected, we enter the second round, but the rule is changed, we choose only two candidates who have the most supporters (if there exist equal votes, select the one has a bigger id for convenience), then vote again, we check if still none have three, vote the two one again. All vote process is random. Since every round may cost resource, so we cannot make the election round and round, never stop. So we want to know if we want have the leader only in two rounds, then which way is better.

To solve the problem, we may model the two ways as below:



(a) The Model of first way (b) The Model of second way

Fig. 4. The models of leader election

Let $D_1=(S_1,P_1,L_1)$ be a DTMCs, described as Figure 4 (a), $D_2=(S_2,P_2,L_2)$ be a DTMCs, described as Figure 4 (b), where $u_1,..,u_j$ are the states representing the voting situation about unable to get a leader, $s_1,..,s_i$ means getting a leader successfully. Since the vote is random, the probability of arriving at u or s is $1/4^4$, the same way for $u_{1'},\ldots,u_{j'}$, $s_{1'}\ldots s_{i'}$ with $1/4^2$. We will simplify them by calculating the total probability

of leader elected on the first round and second round, and combine the states .

Leader must have at least three tickets, so on the first round, the probability of leader elected can calculate as

$$\frac{4 \times C_4^4 + 4 \times C_4^3 \times C_3^1}{4^4} = \frac{13}{64}$$

while none elected on the first round

$$1 - \frac{13}{64} = \frac{51}{64}$$

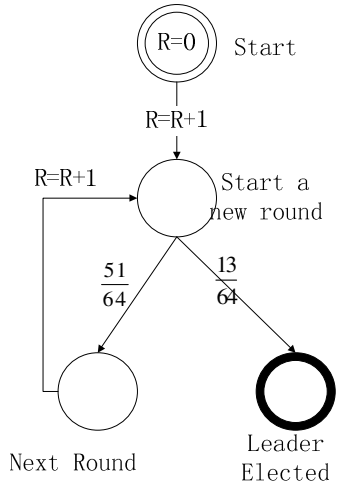
On the second round, while the second way only has two candidates left, the probability can calculate as

$$\frac{2 \times C_4^4 + 2 \times C_4^3}{2^4} = \frac{5}{8}$$

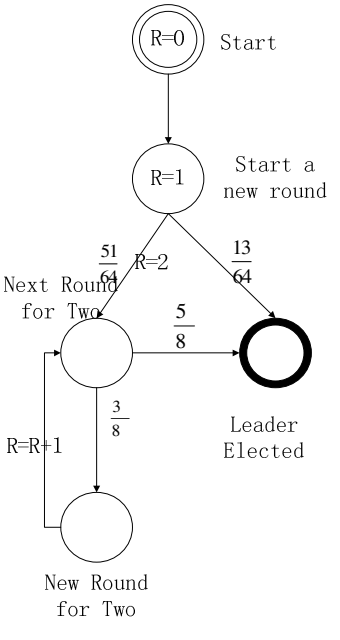
while none elected on the second round

$$1 - \frac{5}{8} = \frac{3}{8}$$

So we simplify both D1 and D2 as below:



(a) The Simplification of Fig 4(a)



(b) The Simplification of Fig 4(b)

Fig. 5. The Simplification of Fig 4 by combining similar states

Then a LTL formula can be used describe the property we concern: we want have the leader only in two rounds. The formula is written as follows:

$$\varphi = \mathrm{F}\,(\mathrm{Leader\_Elected} \wedge \mathrm{R} \leq 2) \tag{6}$$

Now the satisfiability degree of φ on D1 and D2 can be computed

$$\mathrm{f}(\varphi|\mathrm{D}_1) \cong 0.371,$$

$$\mathrm{f}(\varphi|\mathrm{D}_2) \cong 0.701.$$

Obviously, $\mathrm{f}(\varphi|\mathrm{D}_2) > f(\varphi|\mathrm{D}_2)$, that is, though both D1 and D2 have the ability to satisfy φ true, but D2 is more stronger than D1, so we want to have the leader only in two rounds, the second way is better than the first, which is told by their satisfiability degrees.

5 Conclusions

In this paper, the conception of satisfiability degree is introduced into model checking, which is extend from based on propositional Logic to the higher-order logic mainly LTL through the transition system. We not only give the definition of satisfiability degree on transition system but the algorithm of computation on the systems including FPTS and DTMCs. This algorithm modifies the DFS algorithm so that satisfiability degree can be calculated easily. The application of satisfiability degree on transition system is also shown to check which system can satisfy a property better by the example leader election. We show the way how model checking can do quantitative analysis and fault-tolerant analysis for sometimes we must compromise to the real world if we cannot have the certain result, and then the satisfiability degree is a good choice.

References

[1] Luo, G.M., Yin, C.Y., Hu, P.: An algorithm for calculating the satisfiability degree. In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2009), vol. 7, pp. 322–326 (2009)

[2] Yin, C.Y., Luo, G.M., Hu, P.: Backtracking search algorithm for satisfiability degree calculation. In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2009), vol. 2, pp. 3–7 (2009)

[3] Hu, P., Luo, G.M., Yin, C.Y.: Computation of satisfiability degree based on CNF. In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2009), vol. 6, pp. 142–146 (2009)

[4] Clarke, E.M., Orna, G., Peled, D.A.: Model Checking, pp. 13–24. The MIT Press, Cambridge (1999)

[5] Huth, M.l., Ryan, M.: Logic in Computer Science Modeling and Reasoning about Systems, 2nd edn., pp. 172–186 (2005)

[6] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Intruduction to algorithm, 2nd edn., pp. 540–549 (2002)

[7] Itai, A., Rodeh, M.: Symmetry Breaking in Distributed Networks. Information and Computation 88(1), 60–87 (1990)

[8] Hong, I., Sohn, S.H., Lee, J.K., et al.: DFS algorithm with ranking based on genetic algorithm in UHF portable system. In: 2009 9th International Symposium on Communications and Information Technology, ISCIT 2009, pp. 454–459 (2009)

Research on Optimization Design of Bottom Angle of Drag-Reducing Structure on the Riblets Surface

Qi-feng Zhu, Bao-wei Song, and Peng Wang

Northwestern Polytechnical University, Xi'an, 710072, China
zhuqifeng0811@163.com

Abstract. Research on the drag-reducing structure of the haploid interval riblets surface is carried out with the numerical simulation, and in view of its flow field features, this paper appropriately deals with the computing domain, grids and flow parameters during the numerical calculation. By the adoption of iSIGHT software, structure parameters generated program of riblets surface, hydrodynamic numerical simulation software GAMBIT and FLUENT is integrated. With optimized design technology and the numerical simulation technology, the optimization design of the bottom angle on the riblets surface is made by adopting the two-stage combining optimization method which includes overall search based on adaptive simulated annealing algorithm and partial optimal search of sequential quadratic programming. The results show that after the optimization design, the relative amount of drag reduction of the riblets surface reaches 3.506 %. Compared with the initial value, the relative amount of drag reduction increases and the drag-reducing effect is remarkable. This provides a new direction for the research on drag reduction of riblets surface.

Keywords: riblets surface, numerical simulation, drag-reducing effect, optimization design, iSIGHT.

1 Introduction

Widely available in the fluid transport, aviation and maritime areas, turbulent flow drag is the main energy source of the pipeline, conventional aircraft, vessels, and the working parts which frequently contacts the viscous fluid such as submarines, the wings, etc. Drag-reducing of the riblets surface is a turbulent drag reduction technology, which would reach quite good drag reduction effect through making a certain size of the riblets surface on the outer surface[1-2]. This technology has been practically applied abroad, for example, in the Summer Olympic Games of 1984, the U.S. rowing team successfully applied films with a riblets surface structure at the bottom of the row boats; 70% of the testing Airbus A320 surface were covered with the riblets surface structured film, which saved oil by1% ~ 2%; the tests on Learjet aircraft in the NASA Langley Center showed that the drag reduction is about 6% by using this structured film.

Symmetric riblets surface has better drag reduction effect, and this has aroused great concerns of scholars home and abroad, who then carried out a large number of

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 33–40, 2010.

pilot studies and numerical simulation studies. They have done a lot of valuable explorations on the riblets surface shape parameters (height, spacing, and angle) and the turbulent characteristics, burst intensity, frequency and band structure of a turbulent boundary layer, and etc, some scholars study on the optimization of the riblets surface [3-8]. This paper feasibly integrates the numerical simulation with the optimization design, carries out the applied research on drag-reducing structure optimization design with numerical simulation, and applies to optimized design of the bottom angles, so as to verify the feasibility of this method in the optimal design of the bottom angle on the riblets surface. This provides a new direction for the design and research on the drag reduction of riblets surface.

2 Numerical Simulations

2.1 Riblets Surface Structure

With the choosing subject of the symmetric V-shaped riblets surface structure which has been confirmed with the best drag-reducing effect, this paper mainly studies the impact of the size of the bottom angle to the drag-reducing effect of the surface, while the width of riblets surface interval is same with that of the riblets surface. The specific geometric parameters and the structure shape of the riblets surface are shown in Figure 1. The basic size parameters are: s-the width of the riblets surface, α-bottom angle of riblets surface, a-riblets surface interval.

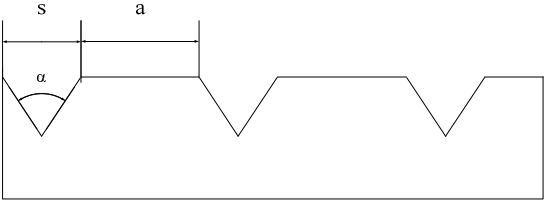


Fig. 1. Riblets Surface Structure

2.2 Control Equations

This paper adopts Reynolds-averaged Navier-Stokes equations and RNG $k-\varepsilon$ turbulence model. This calculating method can provide more precise numerical solution of the flow field, but with large calculating amount. The specific mathematical model [9-10] of this method is as follows:

Continuity equation:

$$\frac{\partial \overline{u_i}}{\partial x_i} = 0 \tag{1}$$

Reynolds-averaged Navier-Stokes equation:

$$\rho \overline{u_j} \frac{\partial \overline{u_i}}{\partial x_j} = \rho \overline{f_i} - \frac{\partial \overline{p}}{\partial x_i} + \frac{\partial}{\partial x_j}(\mu \frac{\partial \overline{u_i}}{\partial x_j} - \rho \overline{u_i u_j}) \tag{2}$$

Turbulent fluctuation kinetic energy equation (k equation) as follows:

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_i}(\rho k u_i) = \frac{\partial}{\partial x_i}\left[\left(\mu_0 + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] + P_k - \rho\varepsilon \tag{3}$$

Turbulent energy dissipation rate (ε equation) as follows:

$$\frac{\partial}{\partial t}(\rho \varepsilon) + \frac{\partial}{\partial x_i}(\rho \varepsilon u_i) = \frac{\partial}{\partial x_j}\left[\left(\mu_0 + \frac{\mu_t}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_j}\right] + C_{\varepsilon 1}\frac{\varepsilon}{k}P_k - C_{\varepsilon 2}\rho\frac{\varepsilon^2}{k} - R_\varepsilon \tag{4}$$

2.3 Calculation Domain and Grid Generation

Taking the computer's ability into account, after repeated experiments, the paper selects a two-dimensional space with riblets flowed surface as the computational domain, the width of the rectangular computational domain is Ly=100*s. the riblets surface length is L=20*s. In order to avoid the effects of the entrance and the exit, a flat section 30 times size of the riblets surface is retained in both entrance and exit. Figure 2 is the final calculation domain.

In this paper, the computational grid is generated by the method of limited volume discretion. In view of its special shape and the existence of cusps on riblets surface, this paper uses unstructured triangular grid to fill [11]. In order to improve the calculation accuracy and save computing resources, the defined size function method is applied in the near wall region to make the grid around the riblets surface turn intensive and gradually become sparse outward. Figure 3 is the partial enlarged drawings of grid in 60° angle around the riblets surface.

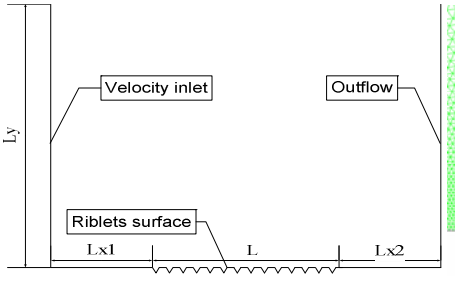


Fig. 2. Computational Domain

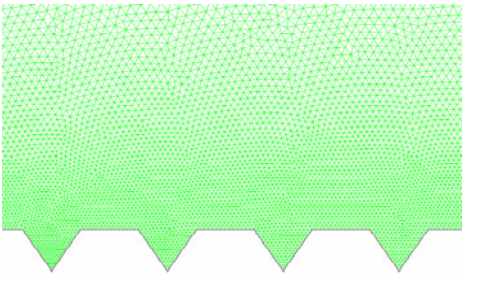


Fig. 3. Partial Enlarged Drawings of Grid

2.4 Solution Conditions

The specific calculation conditions are set as follows:

Solver: steady-state segregated two-dimensional implicit solver.

Turbulence equations: RNG $k-\varepsilon$ turbulence model.

Fluid medium: water, density of 998.2kg/m^3, dynamic viscosity coefficient is 0.001003Pa•s.

Boundary conditions: the entrance is velocity inlet with flow rate 5m/s, along the x-axis; the exit is outflow.

Discrete equation method: pressure-correction method uses Numerical SIMPLEC calculation method [10], and the discrete method of parameters uses the second-order accuracy upwind scheme.

Convergence criteria: the continuity equation residual is set to 1e-4 and 1e-5 for others.

3 Optimization Design

3.1 Model of Optimization Design

By comparing the drag of riblets surface and the corresponding smooth surface, this paper evaluates the drag-reducing effect by relative drag-reducing amount. The total drag of the riblets surface contains viscous drag and pressure drag, while for the smooth surface; it only contains viscous drag, and no pressure drag. The relative amount of drag reduction is defined as follows:

$$d = \frac{F - F_1}{F} \times 100\% \tag{5}$$

F -the total drag of smooth surface; F_1 -the total drag of riblets surface.

The relative amount of drag reduction can clearly express the drag-reducing effect of the riblets surface, Therefore the relative amount of drag reduction is selected as the objective function. As this paper mainly studies how the size of bottom angle of the riblets surface impact the riblets surface, the selected variable is the bottom angle (unit:°) of the riblets surface structure. With comprehensive consideration, the numerical range of the variables is 30 ~ 150, the eventual mathematical model of the optimal design is as follows:

$$\begin{aligned} &Max \quad d = f(s, \alpha, a) \\ &S.t. \quad 30 \le \alpha \le 150 \\ &\qquad s = a = 0.01 \end{aligned} \tag{6}$$

3.2 Optimization Process

Per the requirements, the generated program GoucaoCreat.exe of the ridge-like surface structure parameters is made. Introduce the file shape.txt into GAMBIT, set the model, generate and output the Grid file. FLUENT reads the grid file, gets the solution through setting the parameters. iSIGHT analyzes the hydrodynamics report file trans.out generated by solution got from FLUENT, and outputs the total drag of riblets surface. Through comparison with the total drag of the smooth surface, the relative amount of drag reduction is calculated out and output as the objective function. The process diagram of the GoucaoCreat.exe program, grid division software GAMBIT and FLUENT integrated by iSIGHT is shown in Figure 4.

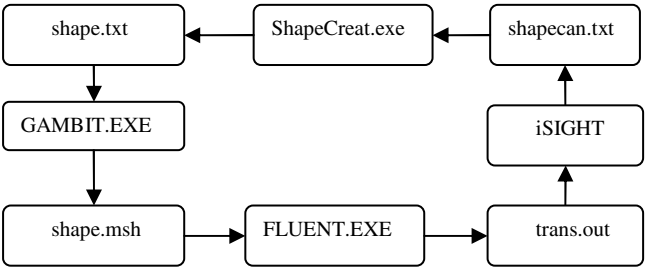


Fig. 4. Optimization Process Chart

3.3 Optimized Algorithm

With the problem that the optimized design of bottom angle on the riblets surface has only one design variable, in order to avoid sinking in the plight of the partial optimized solution, a combined optimized program is developed for the problem. This program is carried out in two steps.

Step one: optimization algorithm more suitable for the overall search is adopted to facilitate the search in the overall design space and initially determine the optimal design point. This paper applies adaptive simulated annealing algorithm (ASA) as the first step of the optimization algorithm.

Step two: Adopting the numerical optimization techniques, one shortcoming of the adaptive simulated annealing algorithm (ASA) is lack of accuracy. In order to solve the problem, a numerical optimization algorithm is chosen to be combined with the global search optimization algorithm. In this paper, the Sequential Quadratic Programming - NLPQL is the optimization method in second step.

4 Results Analysis

With optimized design of variables through the combined optimization design program of ASA + NLPQL, and the 332 steps of iterative computation, this paper ultimately gets the optimal bottom angle of the riblets surface structure.

Figure 5 and Figure 6 are the optimized historical maps of design variables (riblets surface angle α) and the objective function (the relative amount of drag reduction d) VS Run Counter.

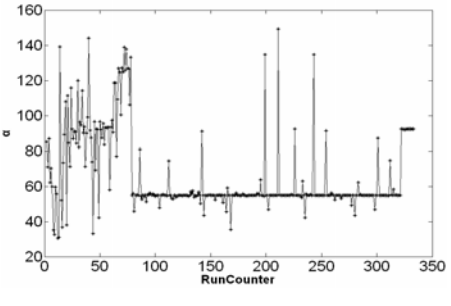


Fig. 5. Optimized Historical Chart of Bottom Angle

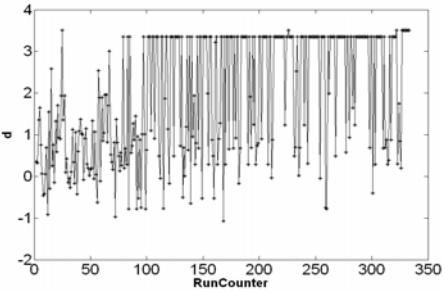


Fig. 6. Optimized Historical Chart of Relative Amount of Drag Reduction

From Fig. 5 and Fig. 6, the following can be getting: during the optimal design process, first, use the adaptive simulated annealing algorithm (ASA) to search the optimal value in the solution range. At this time, parameters changes within a wider range. After determining the initial optimal solution, the sequential quadratic programming algorithm is carried out in the second step to find the precise optimal solution within the partial optimization vicinity, and the overall optimal solution is ultimately obtained.

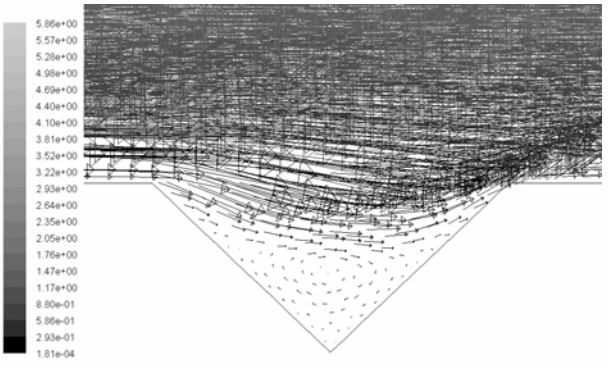


Fig. 7. Velocity Vector Distribution near Bottom Angle

From Figure 7, it can be seen as follows: After the optimization design, a stable low-speed vortex was formatted; whose up flow direction is the same with the free flow. The vortex affects as a roller bearing, and it makes the friction between the free flow and riblets surface reduce, which achieves a drag-reducing purposes.

Table 1 is the contrast of the bottom angle of ridge-shaped surface and the corresponding drag characteristics before and after the optimal design.

From the table above, it can be seen that with optimal design, the bottom angle of the riblets surface becomes large, the total drag reduces, the relative amount of drag reduction increases, reaching 3.506%, and a very remarkable drag reduction effect is achieved.

Table 1. Contrast of Relevant Parameters before and after Optimal Design

Optimal way	Parameters				
	Angle of ridge surface α(°)	Force of the smooth surface (N)	Force of the ridge surface (N)	Relative reduction (%)	Run counter
Initial value	60	0.09492	0.09485	0.0737	
Optimal value	92.3		0.09159	3.5067	332

Table2 is the contrast of the drag characteristics between the riblets surface and smooth surface before and after the optimal design.

Table 2. Contrast of Drag Characteristics between Riblets Surface and Smooth Surface before and after Optimal Design

Optimal way	Parameters					
	Drag of the smooth surface (N)			Drag of the ridges surface (N)		
	Pressure drag	Viscous drag	total drag	Pressure drag	Viscous drag	Total drag
Initial value	0	0.09492	0.09492	0.03281	0.06204	0.09485
Optimal value				0.02470	0.06690	0.09159

During the Optimization design process, the computer constantly changes the bottom angle of ridge-shaped surface based on the optimization program so as to change the flow field characteristics of the surface, thus make the pressure drag and turbulence drag of the riblets surface change. As can be seen from Table 2, compared with the smooth surface, the riblets surface produces the pressure drag. However, the viscous drag of the riblets surface is significantly reduced, and thus causes the decrease of the total drag, reaching the drag reduction effect. In addition, before and post the optimal design, the pressure drag and viscous drag of the riblets surface reduce simultaneously, and the total drag decreases, achieving the optimal design of drag-reducing.

5 Conclusions

This paper puts forward the optimized design program of the bottom angle drag-reducing structure on the riblets surface based on iSIGHT software. Making use of software integration technology, this system integrates the numerical simulation technology and optimized design technology to make optimized design for the bottom angle drag-reducing structure on the riblets surface. In the optimized design process, considering the specificity of the problems, a combined optimization program is developed. Through the comparison of the results before and after optimized design, it can be find that after optimized design, the relative amount of drag reduction of the riblets surface reaches over 3.5067 %. Compared with the initial value, drag reduction

effect is obvious, and achieve the optimal design of the bottom angle on the ridge-shaped surface. In addition, by analyzing the changing characteristics of the pressure drag and viscous drag in the design process, the conclusions can be got as follows:

(1) The existence of the riblets surface changes the flow field characteristic of near-wall regions on the flat area, which resulting in the pressure drag and greatly reduces the viscous drag at the same time. The increase of pressure drag is not sufficient to make up for the decrease of viscous drag, therefore, the total drag reduces and the drag-reducing effect is achieved.

(2) With the optimized design of bottom angle on the riblets surface, the pressure drag and the viscous drag reduce simultaneously, then the total drag decreases, and the relative amount of drag reduction increases, achieving a remarkable drag-reducing effect.

References

1. Chang, Y.F., Jiang, N.: Experimental Study on Coherent Structure Passive Control and Drag Reduction in Turbulent Boundary Layer with Grooved Surface. Journal of Aerospace Power 23, 788–795 (2008)
2. Pan, G., Guo, X.J., Hu, H.B.: Numerical Simulation of Semicircular Traveling Wave Surface and Study on Its Drag-Reduction Mechanism. Journal of System Simulation 18, 3073–3094 (2006)
3. Walsh, M.J.: Riblets as A Viscous Drag Reduction Technique. AIAA Journal 21, 485–486 (1983)
4. Walsh, M.J.: Turbulent Boundary Layer Drag Reduction Using Riblets. AIAA Paper 82-0169 (1982)
5. Walsh, M.J., Lindemann, A.M.: Optimization and Application of Riblets for Turbulent Drag Reduction. AIAA Paper: 84-0347 (1984)
6. Bacher, E.V., Smith, C.R.: A Combine Visualization-Anemometry Study of The Turbulent Drag Reducing Mech-Anisms of Triangular Micro-Groove Surface Modifications. AIAA Paper 85-0548 (1985)
7. Vukoslavcevic, P., Wallance, J.M., Balint, J.L.: Viscous Drag Reduction Using Stream Wise Aligned Riblets. AIAA Journal 30, 1119–1122 (1992)
8. Wu, Y.F., Long, J., Zhang, X.L.: Research on the Optimum Angle of the V-shaped Riblet in the Channel. Journal of Sichuan University of Science and Technology 23, 80–81 (2004)
9. Zhang, Z.X., Dong, Z.N.: Viscosity Hydrodynamics. Tsinghua University Press, Beijing (1998)
10. Wang, F.J.: Computational Fluid Dynamics Analysis. Tsinghua University Press, Beijing (2004)
11. Hu, H.B., Song, B.W., Du, X.X., Ma, J.: Research on Numerical Computation of Near-wall Fluid Field on Grooved Surface. Fire Control and Command Control 33, 54–56 (2008)

Towards Analysis of Semi-Markov Decision Processes

Taolue Chen[1,⋆] and Jian Lu[2,⋆⋆]

[1] FMT, University of Twente, The Netherlands
[2] State Key Laboratory of Novel Software Technology, Nanjing University, China

Abstract. We investigate Semi-Markov Decision Processes (SMDPs). Two problems are studied, namely, the time-bounded reachability problem and the long-run average fraction of time problem. The former aims to compute the maximal (or minimum) probability to reach a certain set of states within a given time bound. We obtain a Bellman equation to characterize the maximal time-bounded reachability probability, and suggest two approaches to solve it based on discretization and randomized techniques respectively. The latter aims to compute the maximal (or minimum) average amount of time spent in a given set of states during the long run. We exploit a graph-theoretic decomposition of the given SMDP based on maximal end components and reduce it to linear programming problems.

1 Introduction

Markov decision processes (MDPs, [9]) provide a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision maker. In this paper, we consider *Semi-Markov Decision Processes* (SMDPs, [7]), which generalize MDPs by (1) allowing the decision maker to choose actions whenever the system state changes; (2) modeling the system evolution in *continuous* time; and (3) allowing time spent in a particular state to follow an arbitrary probability distribution [9].

This paper deals with the analysis of SMDPs. Here, we investigate two important criteria, which unfortunately received relatively scant attentions so far. They are *time-bounded reachability probabilities* and *long-run average fraction of time*. We elaborate them as follows:

– The first problem is referred to as the *time-bounded reachability problem*. Namely, given an SMDP, the aim is to compute the maximal (or minimum) probability to reach – under a given class of schedulers – a certain set of state G within T time units. To solve this problem, we propose a Bellman equation to characterize the maximal time-bounded reachability probability, and suggest two approaches to solve it, based on the discretization and the randomized technique respectively. The dual minimum probability problem can be solved accordingly.

⋆ This author is partially supported by the SRO DSN project of CTIT, University of Twente, and an oversea research funding from the State Key Laboratory of Novel Software Technology.
⋆⋆ This author is partially supported by the Chinese national 973 program (2009CB320702), NSFC (60736015, 60721002), and JSNSF (BK2008017).

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 41–48, 2010.

– The second problem is referred to as the *long-run average fraction of time problem*. Namely, given an SMDP, the aim is to compute the maximal (or minimum) average amount of time – under a given class of schedulers – spent in a given set of states B during the long run of the SMDP. To solve this problem, we exploit a graph-theoretic decomposition of the given SMDP based on the notion of maximal end components and thus reduce it to linear programming problems. The dual minimum probability problem can be also solved along the same vein.

Related works. Time-bounded reachability problem has been tackled in [2] and recently in [8] for continuous-time MDPs; a very similar approach has been applied to model checking continuous stochastic logic (CSL, [1]) for a closely related model akin to CTMDP, i.e., interactive Markov chains (IMCs) in [12]. In this paper, we generalize this work to SMDPs. Long-run average fraction of time problem can be considered as a special case of average reward problem, which has been studied extensively; see e.g. [6,9]. However, most of studies focus on a unichain model, which is rather restrictive. In this paper, we extend this to a general SMDP.

2 Preliminarily

Given a set H, let $\Pr : \mathcal{F}(H) \to [0,1]$ be a probability measure on the *measurable space* $(H, \mathcal{F}(H))$, where $\mathcal{F}(H)$ is a *σ-algebra* (a.k.a. *σ-field*) over H. A semi-Markov decision process is given by:

Definition 1 (SMDP). *A* semi-Markov decision process *(SMDP) $\mathcal{M}$ is a tuple* $(S, s_0, A, \mathbf{P}, \mathbf{Q})$*, where*

– *S is a* finite *set of states, with $s_0 \in S$ the initial state;*
– *A is a* finite *set of actions;*
– *$\mathbf{P} : S \times A \times S \to [0,1]$ is the transition probability function, satisfying that for each $s \in S$ and $\alpha \in A$, $\sum_{s' \in S} \mathbf{P}(s, \alpha, s') \in \{0, 1\}$.*
– *$\mathbf{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ is the continuous probability distribution function. We use $\mathbf{Q}(s, a, s', dt)$ to represent a time-differential.*

Typically, $\mathbf{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ returns a *cumulative distribution function* (cdf), such that $\mathbf{P}(s, \alpha, s') = 0$ implies that $\mathbf{Q}(s, \alpha, s')(t) = 1$. We write $\mathcal{Q} : S \times A \times S \to (\mathbb{R}_{\geq 0} \to [0,1])$ for the corresponding *probability density function* (pdf). Namely, $\mathbf{Q}(s, \alpha, s')(t) = \int_0^t \mathcal{Q}(s, \alpha, s')(\tau) d\tau$; or $\mathbf{Q}(s, \alpha, s', dt) = \mathcal{Q}(s, \alpha, s')(\tau) d\tau$. Usually A is ranged over by $\alpha, \beta, \ldots$. With a bit abuse of notations, for each state s, we write $A(s) = \{\alpha \in A \mid \mathbf{P}(s, \alpha, s') > 0 \text{ for some } s'\}$, i.e., the set of actions which are enabled from s. Clearly, $\mathbf{P}(s, \alpha, s') = 0$ for any s' if $\alpha \notin A(s)$. We assume w.l.o.g. that only internal nondeterminism is allowed, i.e., the actions enabled at each state are pairwise different.

The distribution function $\mathbf{H}$ of state s, defined by $\mathbf{H}(t \mid s, \alpha) = \sum_{s' \in S} \mathbf{P}(s, \alpha, s') \cdot \mathbf{Q}(s, \alpha, s')(t)$, denotes the total holding time distribution in s under the action α, regardless of which successor is selected. To avoid the possibility of an infinite number of decision epochs within finite time (i.e., the zeno behavior), we impose the following assumption, similar to [9]: There exists $\varepsilon > 0$ and $\delta > 0$ such that $\mathbf{H}(\delta \mid s, \alpha) \leq 1 - \varepsilon$, for all $s \in S$ and $\alpha \in A$.

2.1 Semantics

As a stochastic model, in general one is interested in certain events in SMDPs. To measure the probability of events in an SMDP, we use a *path* to represent a single outcome of the associated random experiment. In the continuous-time setting, *timed* paths capture the sojourn times in each state which describe the complete trajectory of the SMDP. They are introduced below. (In the reminder of this paper, paths refer to timed paths.)

Given an SMDP $\mathcal{M}$, $Paths^n(\mathcal{M}) = S \times (A \times \mathbb{R}_{\geq 0} \times S)^n$ is the set of paths of length n in $\mathcal{M}$; the set of *finite* paths in $\mathcal{M}$ is defined as $Paths^\star(\mathcal{M}) = \bigcup_{n\in\mathbb{N}} Paths^n$, and $Paths^\omega(\mathcal{M}) = (S \times A \times \mathbb{R}_{\geq 0})^\omega$ is the set of *infinite* paths in $\mathcal{M}$. Accordingly, $Paths(\mathcal{M}) = Paths^\star(\mathcal{M}) \cup Paths^\omega(\mathcal{M})$ denotes the set of all paths in $\mathcal{M}$. When $\mathcal{M}$ is clear from the context, we denote a path $\sigma \in Paths(s_0)$ if σ starts in state s_0. The same applies to $Paths^n(s_0)$, $Paths^\omega(s_0)$ and $Paths^\star(s_0)$, respectively. Typically, a single path (of length n) σ is denoted $s_0 \xrightarrow{\alpha_0,t_0} s_1 \xrightarrow{\alpha_1,t_1} \cdots \xrightarrow{\alpha_{n-1},t_{n-1}} s_n$, where $|\sigma| = n$ is the length of σ and $\sigma[\downarrow] = s_n$ is the last state of σ. For $k \leq |\sigma|$, $\sigma[k]$ is the $(k+1)$-th state on σ; $\sigma\langle k\rangle := t_k$ is the time spent in state s_k. If $i < j \leq |\sigma|$, then $\sigma[i..j]$ denotes the path-infix $s_i \xrightarrow{\alpha_i,t_i} s_{i+1} \cdots \xrightarrow{\alpha_{j-1},t_{j-1}} s_j$ of σ. Let $\sigma@t$ be the state occupied in σ at time $t \in \mathbb{R}_{\geq 0}$, i.e. $\rho@t := \sigma[n]$ where n is the smallest index such that $\sum_{i=0}^{n} \sigma\langle i\rangle > t$. For infinite path $\sigma = s_0 \xrightarrow{\alpha_0,t_0} s_1 \xrightarrow{\alpha_1,t_1} \cdots$, we require time-divergence, i.e., $\sum_{i\geq 0} t_i = \infty$.

Our goal is to measure the probabilities of (measurable) sets of paths. To this end, we first define a σ-algebra of sets of *combined transitions* [11] which we later use to define σ-algebras of sets of finite and infinite paths. Intuitively, a combined transition is a tuple (α, t, s') which entangles the decision for action α for the time t after which the SMDP moves to successor state s'. Formally, given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, let $\Omega = A \times \mathbb{R}_{\geq 0} \times S$ be the set of combined transitions in $\mathcal{M}$. To define a measurable space on Ω, note that S and A are finite; hence, the corresponding σ-algebras are defined as their power sets respectively. And we use the *Borel σ-field* $\mathcal{B}(\mathbb{R}_{\geq 0})$ to measure the corresponding subsets of $\mathbb{R}_{\geq 0}$. Recall that a Cartesian product is a measurable rectangle if its constituent sets are elements of their respective σ-algebras. We then use $2^A \otimes \mathcal{B}(\mathbb{R}_{\geq 0}) \otimes 2^S$ to denote the set of all measurable rectangles. It generates the desired σ-algebra $\Im = \sigma(2^A \otimes \mathcal{B}(\mathbb{R}_{\geq 0}) \otimes 2^S)$ of sets of combined transitions, which can be used to infer the σ-algebra $\mathcal{F}(Paths^n(\mathcal{M}))$ of sets of paths of length n: $\mathcal{F}(Paths^n(\mathcal{M}))$ is generated by the set of measurable rectangles, that is $\mathcal{F}(Paths^n(\mathcal{M})) = \sigma(\{S_0 \times M_1 \times \cdots \times M_n \mid S_0 \subseteq S, M_i \in \Im\})$. The σ-algebra of sets of *infinite* paths is obtained, in turn, by applying the standard *cylinder* set construction. A set C^n of *paths of length* n is called a cylinder base; it induces the infinite cylinder $C_n = \{\pi \in Paths^\omega(\mathcal{M}) \mid \pi[0..n] \in C^n\}$. A cylinder C_n is measurable if $C^n \in \mathcal{F}(Paths^n(\mathcal{M}))$; C_n is called an infinite rectangle if $C^n = S_0 \times A_0 \times T_0 \times \cdots \times A_{n-1} \times T_{n-1} \times S_n$ and $S_i \subseteq S$, $A_i \subseteq A$ and $T_i \subseteq \mathbb{R}_{\geq 0}$. It is a measurable infinite rectangle, if $S_i \in 2^S$, $A_i \in 2^A$ and $T_i \in \mathcal{B}(\mathbb{R}_{\geq 0})$. We obtain the desired σ-algebra of sets of infinite paths $\mathcal{F}(Paths^\omega(\mathcal{M}))$ as the minimal σ-algebra generated by the set of measurable cylinders.

Schedulers. As in MDPs, the decision making in SMDPs is specified by *schedulers* (or policies). It also defines the semantics of SMDPs, namely, an SMDP and an associated

scheduler induce a unique *probability measure* on the measurable spaces $(Paths^{\omega}(\mathcal{M}), \mathcal{F}(Paths^{\omega}(\mathcal{M})))$. A scheduler quantifies the probability of the next action: If state s is reached via a finite path σ, the scheduler yields a probability distribution over $A(\sigma[\downarrow])$. In the timed setting, the notion of *measurable schedulers* [11] has to be adopted. A measurable scheduler can incorporate the complete information from the history that led into the current state when deciding which of the next actions to take; in particular, it may yield different decisions depending on the time that has passed or in single states.

Definition 2 (Measurable scheduler). *Let $\mathcal{M}$ be an SMDP. A mapping $D : Paths^{\star}(\mathcal{M}) \times A \to [0,1]$ is a* measurable scheduler *if the functions $D(\cdot,\alpha) : Paths^{\star}(\mathcal{M}) \to [0,1]$ are measurable for any $\alpha \in A$.*

We now embark on defining a probability measure on the σ-algebra $\mathcal{F}(Paths^{\omega}(\mathcal{M}))$. To this aim, we first define a probability measure μ_D on the set of combined transitions, i.e., on the measurable space $(\Omega, \Im)$. For all $\sigma \in Paths^{\star}(\mathcal{M})$, $\mu_D(\sigma, M) = \sum_{\alpha \in A} \sum_{s' \in S} D(\sigma,\alpha) \int_{\mathbb{R}_{\geq 0}} \mathbf{1}_M(\alpha,t,s') \mathbf{Q}(s,\alpha,s',dt)$, where $\mathbf{1}_M(\alpha,t,s')$ is the indicator for the set $M \subseteq \Omega$, that is, $\mathbf{1}_M(\alpha,t,s') = 1$ if the combined transition $(\alpha,t,s') \in M$ and 0 otherwise.

With the help of μ_D, we can define the probability measure on the sets of finite paths inductively as follows: $\Pr^0_{s_0,D} : \mathcal{F}(Paths^0(\mathcal{M})) \to [0,1]$ is given as $\Pr^0_{s_0,D}(\Pi) = 1$ if $s_0 \in \Pi$; 0 otherwise. (Note that $Paths^0(\mathcal{M}) \subseteq S$.) And $\Pr^{n+1}_{s_0,D} : \mathcal{F}(Paths^{n+1}(\mathcal{M})) \to [0,1]$ is given as $\Pr^{n+1}_{s_0,D}(\Pi) = \int_{Paths^n(\mathcal{M})} \Pr^n_{s_0,D}(d\pi) \int_{\Omega} \mathbf{1}_{\Pi}(\pi \cdot m) \mu_D(\pi, dm)$. Intuitively, it derives the probability $\Pr^{n+1}_{s_0,D}$ on sets of paths Π of length $n+1$ by multiplying the probability $\Pr^n_{s_0,D}(d\pi)$ of a pat π of length n with the probability $\mu_D(\pi, dm)$ of a combined transition m such that the concatenation $\pi \cdot m$ is a path from the set Π. Hence we obtain measures on all σ-algebras $\mathcal{F}(Paths^n(\mathcal{M}))$ of subsets of paths of length n. This extends to a measure on $(Paths^{\omega}(\mathcal{M}), \mathcal{F}(Paths^{\omega}(\mathcal{M})))$ as follows: First, note that any measurable cylinder can be represented by a base of finite length, i.e., $B_n = \{\sigma \in Paths^{\omega}(\mathcal{M}) \mid \sigma[0..n] \in B^n\}$. Now the measures $\Pr^n_{s_0,D}$ on $\mathcal{F}(Paths^n(\mathcal{M}))$ extend to a unique probability measure $\Pr^{\omega}_{s_0,D}$ on $\mathcal{F}(Paths^{\omega}(\mathcal{M}))$ by defining $\Pr^{\omega}_{s_0,D}(B_n) = \Pr^n_{s_0,D}(B^n)$. The Ionescu-Tulcea extension theorem is applicable due to the inductive definition of the measures $\Pr^n_{s_0,D}$ and assures the extension to be well defined and unique.

3 Time-Bounded Reachability

In this section, we focus on the problem of computing the maximal time-bounded reachability probability in SMDPs. Namely, given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, a set of goal states $G \subseteq S$, a time bound T, we ask: what is the maximal probability to reach G within T? Formally, Let $\Diamond^{[0,T]} G$ denote a set of infinite paths of $\mathcal{M}$ which reach G within time bound T, starting from the initial state s_0. We intend to compute $p_{\max} = \sup_D \{\Pr^{\omega}_{s_0,D}(\Diamond^{[0,T]} G)\}$.

The basic idea to solve this problem is to "encode" the time into the state space, and then to resort to the reachability problem in MDPs. Generally, given a (continuous-state) MDP $(\Xi, A, \mathcal{P})$ with a set of goal states $\Upsilon \in \mathcal{F}(\Xi)$, the maximal reachability

probability can be computed via the following celebrated Bellman equation [3]: $V(\xi) = \max_{\alpha \in A} \left\{ \int_{\Xi} \mathcal{P}((\xi), \alpha, d\xi') \cdot V(\xi') \right\}$ if $\xi \notin \Upsilon$; and 1 otherwise. By instantiating this equation with $\Xi = S \times \mathbb{R}_{\geq 0}$ and $\Upsilon = G \times [0, T]$, we obtain that

- if $s \in G$ and $x \leq T$, $V(s, x) = 1$;
- if $s \in G$ and $x > T$, $V(s, x) = 0$; and
- if $s \notin G$ and $x \leq T$, $V(s, x) =$

$$\begin{aligned} &\max_{\alpha \in A} \left\{ \int_{S \times \mathbb{R}_{\geq 0}} \mathcal{P}((s, x), \alpha, d\xi') \cdot V(\xi') \right\} \\ &= \max_{\alpha \in A} \left\{ \sum_{s' \in S} \int_0^{T-x} \mathbf{P}(s, \alpha, s') \cdot \mathbf{Q}(s, s', \alpha, dt) \cdot V(s', x + t) \right\} \end{aligned} \tag{1}$$

We note that in Eq.(1), the time is encoded in an increasing manner; however, it would be convenient to use countdown, namely, forcing the time x to decrease. In this way, one restricts the state space to $S \times [0, T]$. Moreover, as a convention, we write Eq.(1) in a functional form. We summarize the main result of this section as the following theorem.

Theorem 1. *Let $\mathcal{M} = (S, A, \mathbf{P}, \mathbf{Q})$ be an SMDP and $G \subseteq S$ be a set of goal states. Let $p_{\max} = \sup_D \{\Pr^{\omega}_{s_0, D}(\Diamond^{[0,T]} G)\}$, i.e. the maximal probability to reach G within time bound T, starting from the initial state s_0. Then $p_{\max} = \mathrm{lfp}(\Omega)(s_0, T)$ where $\Omega : (S \times \mathbb{R}_{\geq 0} \to [0, 1]) \to (S \times \mathbb{R}_{\geq 0} \to [0, 1])$ is an operator given as*

$$\Omega(F)(s, z) = \max_{\alpha \in A} \left\{ \sum_{s' \in S} \int_0^z \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', z - t) \right\} \tag{2}$$

and $\mathrm{lfp}(\Omega)$ denotes the least fixpoint of the functional Ω.

Computational methods. Our next task is solve the proposed Bellman equation Eq. (2). A naïve approach is to apply Picard iteration, which is unfortunately, quite inefficient. Instead, we propose two approaches: one is based on discretization; the other one is based on randomization.

Discretization. Assume a step size $h = \frac{1}{N}$ for some $N \in \mathbb{N}$. From Eq. (2), we have

$$\begin{aligned} F(s, T) &= \max_{\alpha \in A} \left\{ \sum_{s' \in S} \int_0^T \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', T - t) \right\} \\ &= \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \cdot F(s', T - t) \right\} \\ &\approx \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \left(\int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \right) \cdot F(s', T - i \cdot h) \right\}. \end{aligned}$$

Hence we can introduce $\tilde{F}$ as an approximation of F:

$$\widetilde{F}(s, T) = \max_{\alpha \in A} \left\{ \sum_{i=0}^{N-1} \sum_{s' \in S} \left(\int_{i \cdot h}^{(i+1) \cdot h} \mathbf{Q}(s, s', \alpha, dt) \cdot \mathbf{P}(s, \alpha, s') \right) \cdot \widetilde{F}(s', T - i \cdot h) \right\},$$

which can be solved by linear programming [3].

Unfortunately, due to the diversity of potential distribution functions used in SMDP, we are not able to derive a plausible error bound between F and $\tilde{F}$, which we left for the future work.

Randomization. Inspired by [10], we proposed a *random* Bellman operator to solve Eq.(2). Suppose $\{x_1, \cdots, x_N\}$ are IID draws with respect to Lebesgue measure λ from the interval $[0, T]$, we define

$$\tilde{\Omega}(F)(s, x_i) = \max_{\alpha \in A} \left\{ \frac{1}{N} \sum_{j=1}^{N} \mathbf{P}(s, \alpha, s') \cdot \mathcal{Q}(s, \alpha, s')(x_j) \cdot F(s', x_j) \right\}. \tag{3}$$

It turns out that Eq.(3) constitutes an approximation of Eq.(2), which is much easier to solve [10]. Note that the operator $\widetilde{\Omega}_N$ is self-approximating: for any function F, one can evaluate $\widetilde{\Omega}_N(F)(s, x)$ at any point $s \in S$ and $x \in [0, T]$ without requiring any explicit interpolation of the values of $\widetilde{\Omega}_N(s, x)$ at the random sample points $x \in \{x_1, \cdots, x_N\}$. In particular, since $\mathcal{Q}$ is a continuous function, $\tilde{\Gamma}_N(V)$ is a (random) continuous function of (s, x) and evaluation of this function at any particular point (s, x) involves nothing more than evaluating the simple formula on the RHS of Eq. (3).

4 Long-Run Average Fraction of Time

In this section, we turn to optimizing another measure, the average amount of time spent in a set of states. Given an SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, we fix a scheduler D, and let σ be a path taken *randomly* from the set $Paths^{\omega}(s_0)$. For $B \subseteq S$, let $\mathbf{1}_B$ denote an indicator with $\mathbf{1}_B(s) = 1$ if $s \in B$ and 0 otherwise. Then the quantity $\mathbf{1}_B(\sigma@x)$ is random variable, indicating whether states in B is occupied at time t when starting in s_0. Based on this, we then define a random variable that cumulates the time spent in some state in B up to time t, starting from state s and normalize it by the time t in order to obtain a measure of the *fraction* of time spent in states form B up to time t, namely, $\mathrm{avg}^{\mathcal{M}}_{B,t}(\sigma) = \frac{1}{t}\int_0^t \mathbf{1}_B(\sigma@x)dx$. Since $\mathrm{avg}^{\mathcal{M}}_{B,t}$ is still a random variable, we can derive its expectation given the scheduler D (which results in $\Pr^{\omega}_{s_0,D}$). This value corresponds to the average fraction of time spent in states from B in the time frame up to t. For the long-run average faction of time, we consider the limit $t \to \infty$, as

$$\lim_{t\to\infty} \mathbb{E}(\mathrm{avg}^{\mathcal{M}}_{B,t}) = \lim_{t\to\infty} \int_{Paths^{\omega}(s_0)} \mathrm{avg}^{\mathcal{M}}_{B,t}(\pi) \overset{\omega}{\Pr}_{s_0,D}(d\pi). \tag{4}$$

We want to maximize this quantity w.r.t. measurable schedulers. Namely, to compute $p_{\max} = \sup_D \lim_{t\to\infty} \mathbb{E}(\mathrm{avg}^{\mathcal{M}}_{B,t})$. The rest of this section is devoted to solving this problem. Some standard notions, in particular, the (maximal) *end component*, can be found in [5]. Intuitively, they represent sets of state-action pairs (below, each one is denoted by (C, ∇) where $C \subseteq S$ and $\nabla \subseteq A$) that once entered, can be followed forever if the scheduler chooses the actions in an appropriate way. Namely, for any scheduler D, a behavior will end up with probability 1 in an end component. (See [5] for a formal account.) Given SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, let $\mathrm{MEC}(\mathcal{M})$ be the set of all *maximal end components* (**MECs**) of $\mathcal{M}$. Note that any two MECs are disjoint. One can identify all of these maximal end components via graph-based analysis [5].

Lemma 1. *Given any SMDP* $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$ *and scheduler* D,

$$\lim_{t\to\infty} \mathbb{E}(\mathrm{avg}_{B,t}^{\mathcal{M}}) = \sum_{H\in \mathrm{MEC}(\mathcal{M})} Prob_D(s_0, \Diamond H) \cdot \lim_{t\to\infty} \mathbb{E}(\mathrm{avg}_{B,t}^{H}).$$

In light of Lem.1, we focus on each MEC of $\mathcal{M}$, which corresponds to a *unichain*[1]. As in [9], we define the long-run average fraction of time alternatively as

$$\lim_{n\to\infty} \frac{\mathbb{E}_{s_0}^{D}\left\{\sum_{i=0}^{n}\int_{\tau_i}^{\tau_{i+1}} \mathbf{1}_B(X_i)dt\right\}}{\mathbb{E}_{s_0}^{D}\left\{\sum_{i=0}^{n}\tau_i\right\}}, \tag{5}$$

where X_i and τ_i are random variables, defined as $X_i(\sigma) = s_i$, $\tau_i(\sigma) = t_i$, given an infinite path $\sigma = s_0 \xrightarrow{\alpha_0,t_0} s_1 \xrightarrow{\alpha_1,t_1} \cdots$.

It turns out that in our setting, these two criteria, given in Eq.(4) and Eq.(5) respectively, coincide. (A proof can be found in [9].) For each $s \in S$ and $\alpha \in A$, we proceed to introduce two measures: (1) $r(s,\alpha) = \int_0^\infty \sum_{s'\in S} \mathbf{P}(s,\alpha,s')\left(\int_0^t \mathbf{1}_B(s)du\right)\mathbf{Q}(s,a,s',dt)$, denoting the expected total "reward" between two decision epochs, given that the system occupies state s at the first decision epoch and the decision maker chooses action α; and (2) $y(s,\alpha) = \mathbb{E}[\mathbf{H}(s,\alpha)]$, denoting the expected length of time until the next decision epoch, given that action α is chosen in state s at the current decision epoch. Note that if $s \in B$, then $r(s,\alpha) = \int_0^\infty \sum_{s'\in S} \mathbf{P}(s,\alpha,s')\cdot t\cdot \mathbf{Q}(s,a,s',dt)$, and if $s \notin B$, $r(s,\alpha) = 0$.

We then introduce a Bellman equation to cope with Eq.(5). For each $s \in S$,

$$h_s = \max_{\alpha\in A}\left\{ r(s,a) - \lambda y(s,a) + \sum_{s'\in S} \mathbf{P}(s,a,s')h_{s'} \right\} . \tag{6}$$

It is known [9] that they have at least one solution, and the value of λ at the solutions is what we desire. Eq.(6) can be solved via linear programming by introducing variables $\{\lambda\} \cup \{h_s \mid s \in S\}$, as follows

$$\begin{array}{l} \text{maximize } \lambda \\ \text{s.t. } h_s \le r(s,\alpha) - \lambda y(s,\alpha) + \sum_{s'\in S} \mathbf{P}(s,\alpha,s')h_{s'} \ . \end{array} \tag{7}$$

It turns out [9] that the linear programming given in (7) admits at least one solution, and the variable λ assumes the same value at all solutions.

Now for each MEC H, we can compute the maximum λ_H thanks to the linear programming given in Eq.(7), which paves the way to solve the problem of computing the maximal long-run average fraction of time by reducing to the problem of computing the *maximal expected reachability reward* in an MDP, which is a minor variant of *stochastic shortest path problem* [4]. Let $U = \{s \in H \mid H \in \mathrm{MEC}(\mathcal{M})\}$, i.e., the set of states belonging to some MEC H of $\mathcal{M}$. Since any two MECs are disjoint, for each state

[1] Meaning that for each deterministic stationary scheduler, the resulting discrete-time Markov chain consists of a single recurrent class plus a possibly empty set of transient states.

$s \in U$, we can associate a value λ_H where H is the MEC to which s belongs. We then introduce the following linear programming with a family of variables $\{x_s \mid s \in S \backslash U\}$:

$$\begin{array}{l} \text{minimize } \sum_{s \in S \backslash U} x_s \\ \text{s.t. } x_s \geq \sum_{s' \in S \backslash U} \mathbf{P}(s, \alpha, s') \cdot x_{s'} + \sum_{s' \in U} \mathbf{P}(s, \alpha, s') \cdot \lambda_H. \end{array} \tag{8}$$

x_{s_0} is what we desire.

Algorithm. We summarize the procedure outlined before in Algo. 1.

Algorithm 1

Require: An SMDP $\mathcal{M} = (S, s_0, A, \mathbf{P}, \mathbf{Q})$, a set of states $B \subseteq S$
Ensure: $\sup_D \lim_{t \to \infty} \mathbb{E}(\text{avg}_{B,t}^{\mathcal{M}})$

1: Compute $\text{MEC}(\mathcal{M}) = \{(C_1, \nabla_1), \cdots, (C_n, \nabla_n)\}$;
2: **for** $1 \leq i \leq n$ **do**
3: Compute $r(s, \alpha)$ and $y(s, \alpha)$ for $s \in C_i$ and $\alpha \in \nabla_i(s)$, depending on B;
4: Solve linear programming (7);
5: **end for**
6: Solve linear programming (8);
7: **return** x_{s_0} in (8).

References

1. Baier, C., Haverkort, B.R., Hermanns, H., Katoen, J.-P.: Model-checking algorithms for continuous-time Markov chains. IEEE Trans. Software Eng. 29(6), 524–541 (2003)
2. Baier, C., Hermanns, H., Katoen, J.-P., Haverkort, B.R.: Efficient computation of time-bounded reachability probabilities in uniform continuous-time markov decision processes. Theor. Comput. Sci. 345(1), 2–26 (2005)
3. Bertsekas, D.P.: Dynamic Programming and Optimal Control. Athena Scientific, Belmont (2007)
4. Bertsekas, D.P., Tsitsiklis, J.N.: An analysis of stochastic shortest path problems. Mathematics of Operations Research 16(3), 580–595 (1991)
5. de Alfaro, L.: How to specify and verify the long-run average behavior of probabilistic systems. In: LICS, pp. 454–465 (1998)
6. Guo, X., Hernández-Lerma, O.: Continuous-Time Markov Decision Processes. Springer, Heidelberg (2009)
7. Jewell, W.S.: Markov-renewal programming I: Formulation, finite returen models; markov-renewal programming II: infinite return models, example. Operations Research 11, 938–971 (1963)
8. Neuhaeusser, M.R.: Model Checking Nondeterministic and Randomly Timed Systems. PhD thesis (2010)
9. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
10. Rust, J.: Using randomization to break the curse of dimensionality. Econometrica 65(3), 487–516 (1997)
11. Wolovick, N., Johr, S.: A characterization of meaningful schedulers for continuous-time markov decision processes. In: Asarin, E., Bouyer, P. (eds.) FORMATS 2006. LNCS, vol. 4202, pp. 352–367. Springer, Heidelberg (2006)
12. Zhang, L., Neuhaeusser, M.R.: Model checking interactive markov chains. In: Esparza, J., Majumdar, R. (eds.) TACAS 2010. LNCS, vol. 6015, pp. 53–68. Springer, Heidelberg (2010)

Stability of Equilibrium Solution and Periodical Solution to Cohen-Grossberg Neural Networks

Jingsheng Lei[1], Ping Yan[2,⋆], and Teng Lv[3]

[1] College of Computer and Information Engineering,
Shanghai University of Electronic Power, Shanghai, 200090, P.R. China
[2] School of Science,
Anhui Agricultural University, Hefei, 230036, P.R. China
want2fly2002@163.com
[3] Teaching and Research Section of Computer,
Artillery Academy, Hefei, 230031, P.R. China

Abstract. In this paper, we study delayed reaction-diffusion Cohen-Grossberg neural networks with Dirichlet boundary conditions. By using topology degree theory and constructing suitable Lyapunov functional, some sufficient conditions are given to ensure the existence, uniqueness and globally exponential stability of the equilibrium point. At the same time, another sufficient conditions are also given to ensure the existence and exponential convergence of the periodical solution.

Keywords: Cohen-Grossberg neural networks, reaction diffusion term, globally exponential stability, periodical solution.

1 Introduction

Since Cohen-Grossberg neural networks(CGNNs)which include the traditional neural networks as special cases were first introduced in 1983 by Cohen and Grossberg in [1]. CGNNs have been extensively investigated and successfully applied to parallel computation, associative memory and optimization problems, etc [2–17]. Because these applications heavily depend on the dynamical behaviors of the networks, the analysis of the dynamical behaviors is the necessary step to design of neural networks.

However, both in biological and man-made neural networks, diffusion effect cannot be avoided when electrons are moving in asymmetric electromagnetic field, thus we must consider that the activations vary in space as well as in time. Refs [13,16,18–24] have considered the stability of neural networks with diffusion terms, which are expressed by partial differential equations.

Studies on the dynamical behaviors of the neural networks not only involve a discussion of stability properties, but also involve many dynamic behaviors such as periodic oscillatory behavior, bifurcation and chaos [25–26]. In many applications, the properties of periodic oscillatory solutions to CGNNs are of great

⋆ Corresponding author.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 49–56, 2010.

interest, for example, the human brain is in periodic oscillatory of chaos, hence it is of prime importance to study periodic oscillatory and chaos phenomena of neural networks.

Moreover, because of the finite processing speed of information, it is sometimes necessary to take account of time delays in the modelling of the biological or artificial neural networks. Time delays may led to bifurcation, oscillation, divergence or instability which may be harmful to a system, thus, the study of neural dynamics with consideration of the delayed problem becomes extremely important to manufacture high quality neural networks. Refs [2,4–9,12–23] have studied the stability of delayed neural networks.

To the best of our knowledge, few authors have considered global exponential stability and periodicity of delayed reaction-diffusion CGNNs with Dirichlet boundary conditions. But it is important in theories and applications, and also is a very challenging problem. In this paper, we will investigate global exponential stability and periodicity of a class of delayed reaction-diffusion CGNNs with Dirichlet boundary conditions.

This paper is organized as follows. In section 2, model description and main results are given. In section 3, by employing topology degree theory and constructing suitable Lyapunov functional, some sufficient conditions are obtained to ensure the existence, uniqueness and the global exponential stability of equilibrium solution. At the same time, another sufficient conditions are also given to ensure the existence of the periodic oscillatory solutions. Section 4 is the conclusions of the paper.

2 Model Description and Main Results

Consider the following delayed reaction-diffusion Cohen-Grossberg neural networks with Dirichlet boundary conditions:

$$\begin{aligned}\frac{\partial u_i(x,t)}{\partial t} &= \sum_{k=1}^{m}\frac{\partial}{\partial x_k}(D_{ik}\frac{\partial u_i(x,t)}{\partial x_k}) - a_i(u_i(x,t))\Big[b_i(u_i(x,t)) \\ &\quad - \sum_{j=1}^{n} c_{ij} f_j(u_j(x,t)) - \sum_{j=1}^{n}\omega_{ij} f_j(u_j(x,t-t_{ij})) + I_i\Big], \qquad (1)\\ u_i(x,t) &= \varphi_i(x,t), \quad -\tau \le t \le 0,\\ u_i(x,t) &= 0, \quad x \in \partial\Omega.\end{aligned}$$

Where $i = 1, \cdots, n$. n is the number of neurons in the networks. $u(x,t) = (u_1(x,t), \cdots, u_n(x,t))^T$, $u_i(x,t)$ denotes the state of the ith neural unit at time t and in space $x \in \Omega$. Ω is a bounded open domain in $\mathbb{R}^m$ with smooth boundary $\partial\Omega$ and $mes\Omega > 0$ denotes the measure of Ω. $D_{ik} \ge 0$ corresponds to the transmission diffusion coefficient along the ith unit. $a_i(\cdot)$ and $b_i(\cdot)$ represent an amplification function and an appropriately behaved function, respectively. c_{ij} represents the strength of the jth neuron connecting on the ith neuron at time t and in space x, and ω_{ij} represents the strength of the jth neuron connecting on

the ith neuron at time $t-t_{ij}$ and in space x. t_{ij} corresponds to the transmission delays along the axon of the jth neuron form the ith neuron and satisfies $0 \leq t_{ij} \leq \tau$, where $\tau > 0$ is a given constant. f_i shows how the ith neuron reacts to the input. $I(t) = (I_1(t), \cdots, I_n(t))^T$, I_i is the input from outside the system. $\varphi(x,t) = (\varphi_1(x,t), \cdots, \varphi_n(x,t))^T$, $\varphi_i(x,t)$ are given smooth function defined on $[-\tau, 0] \times \Omega$ with the norm $\|\varphi\|_2 = \sqrt{\sum_{i=1}^n \int_\Omega |\varphi_i(x)|_\tau^2 dx}$, where $|\varphi_i(x)|_\tau = \sup_{-\tau \leq s \leq 0} |\varphi_i(x,s)|$.

Definition 1. $u(x,t) = u^* \in \mathbb{R}^n$ is said to be an equilibrium point of system (1) in which I_i $(i = 1, \cdots, n)$ are constants, if for any $i = 1, \cdots, n$, the constant vector $u^* = (u_1^*, \cdots, u_n^*)^T$ satisfies

$$b_i(u_i^*) - \sum_{j=1}^n c_{ij} f_j(u_j^*) - \sum_{j=1}^n \omega_{ij} f_j(u_j^*) + I_i = 0.$$

Definition 2. The equilibrium point $u(x,t) = u^* \in \mathbb{R}^n$ of system (1) in which I_i $(i = 1, \cdots, n)$ are constants, is said to be global exponential stability, if there exist positive constants $\gamma > 0$ and $\varepsilon > 0$ such that

$$\|u(t) - u^*\| \leq \gamma \|\varphi - u^*\| \exp^{-\varepsilon t}, \quad \forall t \geq 0.$$

Throughout the paper, we always assume that system (1) has a smooth solution $u(t,x)$ with the norm $\|u(t)\|_2 = \sqrt{\sum_{i=1}^n \int_\Omega |u_i(x,t)|^2 dx}$, $\forall t \in [0, +\infty)$.

In this paper, when we consider the existence, uniqueness and the global exponential stability of equilibrium solution to system (1), we always assume that:

$(H1)$ There exist constants Δ_{ik}, m_i, M_i and such that $0 < m_i \leq a_i(u_i) \leq M_i$, $i = 1, \cdots, n$ and $D_{ik} \geq \Delta_{ik} \geq 0, (k = 1, \cdots, m)$, $\Delta_i^{min} = \min_k\{\Delta_{ik}\} \geq 0$.

$(H2)$ $b_i(\cdot)$ is differentiable, $b_i(0) = 0$, $i = 1, \cdots, n$. Moreover, $\alpha_i = \inf_{x \in \mathbb{R}}\{\dot{b}_i(x)\} > 0$, where $\dot{b}_i(\cdot)$ denotes the derivative of $b_i(\cdot)$, $i = 1, \cdots, n$.

$(H3)$ For any given $i = 1, \cdots, n$, there exists positive constant β_i such that

$$|f_i(x_i) - f_i(y_i)| \leq \beta_i |x_i - y_i|, \ \forall x_i, y_i \in \mathbb{R}. \tag{2}$$

$(H4)$ There exist constants $\delta > 0$, p_{ij}, q_{ij}, p_{ij}^*, q_{ij}^*, $(i, j = 1, \cdots, n)$ such that for all $i = 1, \cdots, n$, it holds

$$-2\alpha_i m_i - 2\frac{\Delta_i^{min}}{\delta} + \sum_{j=1}^n \Big[M_i \Big(|c_{ij}|^{2p_{ij}} \beta_j^{2q_{ij}} + |\omega_{ij}|^{2p_{ij}^*} \beta_j^{2q_{ij}^*} \Big) + M_j \Big(|c_{ji}|^{2-2p_{ji}} \beta_i^{2-2q_{ji}} + |\omega_{ji}|^{2-2p_{ji}^*} \beta_i^{2-2q_{ji}^*} \Big) \Big] < 0.$$

One of our main results is the following theorem:

Theorem 1. *If $(H1) - (H4)$ hold and $I_i, (i = 1, \cdots, n)$ are all constants, then system (1) has a unique equilibrium solution u^*. And the equilibrium solution u^* of system (1) is global exponential stability.*

In the next part of this section, we will consider the existence of periodic oscillatory solution to system (1), and we always assume that:

$(HP1)$ There exist constants Δ_{ik}, m_i, M_i and such that $0 < m_1 \leq a_i(u_i) \leq M_i,\ i = 1, \cdots, n$ and $D_{ik} \geq \Delta_{ik} \geq 0, (k = 1, \cdots, m)$, $\Delta_i^{min} = \min_k\{\Delta_{ik}\} \geq 0$. Moreover, $a_i(\cdot)$ is differentiable and there exists positive constants Γ_i such that $0 < a_i'(\cdot) \leq \Gamma_i, (i = 1, \cdots, n)$.

$(HP2)$ $I_i(t)$ is $\omega-$periodic function satisfying $I_i(t+\omega) = I_i(t)$ with the boundary $B_i \geq 0,\ (i = 1, \cdots, n)$. $b_i(\cdot) \geq 0$ is differentiable, $b_i(0) = 0,\ i = 1, \cdots, n$. Moreover, $\alpha_i = \inf_{x\in\mathbb{R}}\{\dot{b}_i(x)\} > 0$, where $\dot{b}_i(\cdot)$ denotes the derivative of $b_i(\cdot)$, $i = 1, \cdots, n$.

$(HP3)$ For any given $i = 1, \cdots, n$, there exist positive constants Λ_i and β_i such that

$$\|f_i\| = \sup_u |f_i(u)| \leq \Lambda_i,\ |f_i(x_i) - f_i(y_i)| \leq \beta_i|x_i - y_i|,\ \forall x_i, y_i \in \mathbb{R}. \tag{3}$$

$(HP4)$ There exist constants $\delta > 0$, p_{ij}, q_{ij}, p_{ij}^*, q_{ij}^*, $(i, j = 1, \cdots, n)$ such that for all $i = 1, \cdots, n$, it holds

$$\begin{aligned}
&-2\alpha_i m_i - 2\frac{\Delta_i^{min}}{\delta} + \sum_{j=1}^{n} \Gamma_i\big(\big(|c_{ij}| + |\omega_{ij}|\big)\Lambda_j + B_i\big) \\
&+ \sum_{j=1}^{n} \Big[M_i\Big(|c_{ij}|^{2p_{ij}}\beta_j^{2q_{ij}} + |\omega_{ij}|^{2p_{ij}^*}\beta_j^{2q_{ij}^*}\Big) \\
&+ M_j\Big(|c_{ji}|^{2-2p_{ji}}\beta_i^{2-2q_{ji}} + |\omega_{ji}|^{2-2p_{ji}^*}\beta_i^{2-2q_{ji}^*}\Big)\Big] < 0.
\end{aligned}$$

Another main result is as follows.

Theorem 2. *If $(HP1) - (HP4)$ hold, then system (1) has only one $\omega-$periodic solution and all other solutions converge exponentially to it as $t \to +\infty$.*

3 The Stability of the Equilibrium Solution and the Periodical Solution

We first discuss that system (1) has a unique equilibrium point u^*.

Lemma 3. *If $(H1)$-$(H4)$ hold, then system (1) has a unique equilibrium point.*

Proof. For $i = 1, \cdots, n$, let

$$h_i(u) = b_i(u_i) - \sum_{j=1}^{n} c_{ij}f_j(u_j) - \sum_{j=1}^{n} \omega_{ij}f_j(u_j) + I_i. \tag{4}$$

And define $H_i(u, \lambda) : \bar{Q} \times [0, 1] \to \mathbb{R}^n$ as follows:

$$H_i(u, \lambda) = b_i(u_i) - \lambda\Big[\sum_{j=1}^{n} c_{ij}f_j(u_j) + \sum_{j=1}^{n} \omega_{ij}f_j(u_j) - I_i\Big],$$

where $\lambda \in [0,1], u \in \bar{Q} \subset \mathbb{R}^n$.

By $(H2)$ and $(H3)$, for any given $i = 1, \cdots, n$, we get

$$|b_i(u_i)| \geq \alpha_i |u_i|, \quad |f_i(u_i)| \leq \beta_i |u_i| + |f_i(0)|. \tag{5}$$

By $(H4)$ and $M_i \geq m_i$, there exist the following constants

$$\delta_0 = \min_i \{\alpha_i M_i - \frac{1}{2} \sum_{j=1}^{n} (M_i |c_{ij}|^{2p_{ij}} \beta_j^{2q_{ij}} + M_j |c_{ji}|^{2-2p_{ji}} \beta_i^{2-2q_{ji}})$$
$$+ \sum_{j=1}^{n} (M_i |\omega_{ij}|^{2p^*_{ij}} \beta_j^{2q^*_{ij}} + M_j |\omega_{ji}|^{2-2p^*_{ji}} \beta_i^{2-2q^*_{ji}})]\} > 0$$

and

$$L_0 = \max_i \Big\{ \Big(\int_\Omega M_i^2 [\sum_{j=1}^{n} (|c_{ij}| + |\omega_{ij}|) |f_j(0)| + |I_i|]^2 dx \Big)^{\frac{1}{2}} \Big\},$$

such that

$$\int_\Omega \sum_{i=1}^{n} M_i |u_i| |H_i((u,\lambda)| dx \geq \delta_0 \|u(t)\|_2^2 - L_0 \|u(t)\|_2.$$

Let $R_0 > \frac{L_0}{\delta_0}$, Let $U(R_0) = \{u(t) \in \mathbb{R}^n | \ \|u(t)\|_2 < R_0, \forall t \geq 0\}$. For any $u(t) \in \partial U(R_0) = \{u(t) | \ \|u(t)\|_2 = R_0, \forall t \geq 0\}$, for all $\lambda \in [0,1]$, we have

$$\int_\Omega \sum_{i=1}^{n} M_i |u_i| |H_i((u,\lambda)| dx \geq \delta_0 (\|u(t)\|_2 - \frac{L_0}{\delta_0}) \|u(t)\|_2 > 0.$$

Hence, we know that For any $u(t) \in \partial U(R_0), \forall \lambda \in [0,1]$, it holds $H(u,\lambda) = (H_1(u,\lambda), \cdots, H_n(u,\lambda))^T \neq 0$.

It is easy to know that there exists at least a $u^* \in U(R_0)$ such that $h(u^*) = 0$. It means that system (1) has at least an equilibrium point $u^* \in U(R_0)$.

From $(H2)$-$(H4)$, it is easy to prove the uniqueness of the equilibrium point u^* of system (1). Then the proof of Lemma 4 is finished.

By taking Laypunov functional as follows:

$$V(t) = \sum_{i=1}^{n} \int_\Omega \Big[|y_i(x,t)|^2 e^{2\lambda t} + \sum_{j=1}^{n} M_i |\omega_{ij}|^{2-2p^*_{ij}} \beta_j^{2-2q^*_{ij}} \int_{t-t_{ij}}^{t} |y_j(x,s)|^2 e^{2\lambda(s+t_{ij})} ds \Big] dx, \tag{6}$$

we can get the global exponential stability of the unique equilibrium point u^* of system (1) as follows.

Lemma 4. *If $(H1)$-$(H4)$ hold, then the unique equilibrium point u^* of system (1) is global exponential stability.*

By Lemma 3 and Lemma 4, we can easily get the proof of Theorem 1.

Now we begin to talk about the periodic oscillatory solution to the system (1) satisfying the assumptions $(HP1)$-$(HP4)$.

For any smooth vector functions φ and ψ, let $u(x,t;\varphi) = (u_1(x,t;\varphi),\cdots, u_n(x,t;\varphi))^T$ and $u(x,t;\psi) = (u_1(x,t;\psi),\cdots,u_n(x,t;\psi))^T$ denote the solutions to system (1) satisfying the assumptions $(HP1)$-$(HP4)$ through $(\varphi,0)$ and $(\psi,0)$, respectively.

Let $v_i(x,t) = u_i(x,t;\varphi) - u_i(x,t;\psi),\ i = 1,\cdots n$, by $(HP4)$, there exists a small positive constant $0 < \lambda < \min\{\frac{1}{2}, \min_i\{m_i\alpha_i + \sum_{k=1}^m \frac{\Delta_i^{min}}{\delta}\}\}$ such that

$$\begin{aligned}\widetilde{W}_i = \lambda - \alpha_i m_i - \frac{\Delta_i^{min}}{\delta} + \frac{1}{2}\Big[\sum_{j=1}^n \Gamma_i\big((|c_{ij}| + |\omega_{ij}|)\Lambda_j + B_i\big) \\ + \sum_{j=1}^n (M_i|c_{ij}|^{2p_{ij}}\beta_j^{2q_{ij}} + M_j|c_{ji}|^{2-2p_{ji}}\beta_i^{2-2q_{ji}}) \\ + \sum_{j=1}^n (M_i|\omega_{ij}|^{2p^*_{ij}}\beta_j^{2q^*_{ij}} + M_j|\omega_{ji}|^{2-2p^*_{ji}}\beta_i^{2-2q^*_{ji}}e^{2\lambda\tau})\Big] \le 0.\end{aligned} \tag{7}$$

Taking Laypunov functional just as (6), calculating $D^+V(t)$ along the system decided by $v_i(x,t)$, noting $(HP1)$-$(HP3)$ we can have

$$D^+V(t) \le 2e^{2\lambda t}\sum_{i=1}^n \widetilde{W}_i \int_\Omega |v_i(x,t)|^2 dx \le 0,$$

where $\widetilde{W}_i$ is defined in (7). Let

$$\widetilde{\Upsilon} = \max_i\Big\{1 + (e^{2\lambda\tau} - 1)\sum_{j=1}^n \frac{M_i}{2\lambda}|\omega_{ij}|^{2-2p^*_{ij}}\beta_j^{2-2q^*_{ij}}\Big\} > 1.$$

For $V(t) \ge \|v(t)\|_2^2 e^{2\lambda t}$, it holds $\|v(t)\|_2^2 \le \widetilde{\Upsilon}\|\varphi - \psi\|_2^2 e^{-2\lambda t}$, that is

$$\|u(\cdot,t;\varphi) - u(\cdot,t;\psi)\|_2 \le \sqrt{\widetilde{\Upsilon}}\|\varphi - \psi\|_2 e^{-\lambda t}. \tag{8}$$

For $\tau > 0$, we can choose a positive integer N and a positive constant κ such that

$$\sqrt{\widetilde{\Upsilon}}e^{-\lambda(N\omega - \tau)} \le \kappa < 1. \tag{9}$$

If we define a poincaré mapping $P:\ C \to C$ (here C denotes the continuous function space) by $P(\Phi) = u(x,\omega+\theta;\Phi),\ (\theta \in [-\tau,0])$, then $P^N(\Phi) = u(x,N\omega+\theta;\Phi),\ \theta \in [-\tau,0]$.

Let $t = N\omega$, we get

$$\|P^N(\varphi) - P^N(\psi)\|_2 = \|u(x,N\omega+\theta;\varphi) - u(x,N\omega+\theta;\psi)\|_2 \le \kappa\|\varphi - \psi\|_2.$$

This implies that P^N is a contraction mapping because of $0 < \kappa < 1$. According to the Banach fixed point theorem, there exists a unique point $\varphi^* \in C$ such that $P^N(\varphi^*) = \varphi^*$. Since $P^N(P(\varphi^*)) = P(P^N(\varphi^*)) = P(\varphi^*)$, then $P(\varphi^*)$ is also a fixed point of P^N. By the uniqueness of the fixed point of P^N, we have

$$P(\varphi^*) = \varphi^* \text{ that is } u(x, \omega + \theta; \varphi^*) = \varphi^*.$$

Let $u(x, \omega+\theta; \varphi^*)$ be the solution of CGNNs (1) through $(\varphi^*, 0)$, then $u(x, t+\omega+\theta; \varphi^*)$ is also the solution of CGNNs (1) through $(u(x, \omega+\theta; \varphi^*), 0)$. obviously, for all $t \geq 0$ and $\theta \in [-\tau, 0]$, it has $u(x, t+\omega+\theta; \varphi^*) = u(x, t+\theta; u(x, \omega+\theta; \varphi^*)) = u(x, t+\theta; \varphi^*)$. Hence, for all $t \geq 0$, it has $u(x, t+\omega; \varphi^*) = u(x, t; \varphi^*)$. This shows that $u(x, t; \varphi^*)$ is exactly one ω-periodic solution of CGNNs (1), and according to (14) it is easy to see that all other solutions of CGNNs (1) converge exponentially to it as $t \to +\infty$. Then we obtain Theorem 2.

4 Conclusion

In this paper, the sufficient conditions that ensure the existence and exponential convergence of the periodic solution are also given. These conditions are useful in the design and applications of Cohen-Grossberg neural networks model with delays and reaction-diffusion terms and the method in this paper may be extended for more complex networks.

Acknowledgment

The research is supported by Natural Science Foundation of China(No.60863001), College Science Research Plan Project of Xinjiang Uighur Autonomous Region (No. XJEDU2007I03 and XJEDU2007G01), and Science Research Foundation of Artillery Academy.

References

1. Cohen, M., Grossberg, S.: Absolute stability and global pattern formation and parallel memory storage by competitive neural networks. IEEE Trans. Syst. Man Cybernet. SMC-13, 15 (1983)
2. Cao, J., Liang, L.: Boundedness and stability for Coheb-Grossberg neural network with time-varying delays. J. Math. Anal. Appl. 296, 665–685 (2004)
3. Cao, J., Li, X.: Stability in delayed Cohen-Grossberg neural networks: LMI optimization approach. Physica D 212, 54–65 (2005)
4. Chen, Z., Zhao, H., Ruan, J.: Dynamic analysis of high-order Cohen-Grossberg neural networks with time delay. Chaos Solitons & Fractals 24, 421–430 (2006)
5. Jiang, H., Cao, J.: Dynamics of Cohen-Grossberg neural networks with time-varying delays. Phys. Lett. A 354, 414–422 (2006)
6. Chen, T., Rong, L.: Robust global exponential stability of Cohen-Grossberg neural networks with time delays. IEEE Trans. Neur. Netw. 15, 203–206 (2004)

7. Li, Y.: Existence and stability of periodic solutions for Cohen-Grossberg neural networks with multiple delays. Chaos Solitons & Fractals 20, 459–466 (2004)
8. Hwang, C., Cheng, C., Li, T.: Globally exponential stability of generalized Cohen-Grossberg neural networks with time delay. Phys. Lett. A 319, 157–166 (2003)
9. Wan, H., Qiao, H., Peng, J.: Delay-independent criteria for exponential stability generalized Cohen-Grossberg neural networks with dicrete delay. Phys. Lett. A 353, 151–157 (2006)
10. Liao, X., Li, C., Wong, K.: Criteria for exponential stability of Cohen-Grossberg neural networks. Neural Networks 17, 1401–1414 (2004)
11. Xiong, W., Xu, D.: Global exponential stability of discrete-time Cohen-Grossberg neural networks. Neurocomputing 64, 433–446 (2005)
12. Xu, J., Liu, Z., Liao, X.: Global asymptotic stability of high-order Hopfield type neural networks with time delays. Comp. Math. Appl. 45, 1729–1737 (2003)
13. Yan, P., Lv, T.: Stability of delayed reaction diffusion high-order Cohen-Grossberg neural networks with variable coefficient. In: Proccedings of Second International Symposium on Intelligent Information Technology Application, vol. 1, pp. 741–745. IEEE Computer Society Press, Los Alamitos (2008)
14. Yuan, K., Cao, J.: Global exponential stability of Cohen-Grossberg neural networks with multiple time-varying delays. In: Yin, F.-L., Wang, J., Guo, C. (eds.) ISNN 2004. LNCS, vol. 3173, pp. 745–753. Springer, Heidelberg (2004)
15. Zhang, J., Suda, Y., Komine, H.: Global exponential Stability of Cohen-Grossberg neural networks with variable delays. Phys. Lett. A 338, 44–50 (2005)
16. Zhao, H., Wang, K.: Dynamical behaviors of Cohen-Grossberg neural networks with delays and reaction-diffusion terms. Neurocomputing 70, 536–543 (2006)
17. Chen, T., Rong, L.: Delay-independent stability analysis of Cohen-Grossberg neural networks. Phys. Lett. A 317, 436–449 (2003)
18. Cao, J., Liang, L., Lamb, J.: Exponential Stability of high-order bidirectional associative memory neural networks with time delays. Physica D 199, 425–436 (2004)
19. Liang, J., Cao, J.: Global exponential Stability of reaction-diffusion recurrent neural networks with time-varying delays. Phys. Lett. A 314, 434–442 (2003)
20. Song, Q., Cao, J.: Global exponential Stability and existence of periodic solutions in BAM neural networks with distributed delays and reaction-diffusion terms. Chaos Solitons & Fractals 23, 421–430 (2005)
21. Song, Q., Zhao, Z., Li, Y.: Global exponential Stability of BAM neural networks with distributed delays and reaction-diffusion terms. Phys. Lett. A 335, 213–225 (2005)
22. Song, Q., Cao, J., Zhao, Z.: Periodic solutions and its exponential stability of reaction-diffusion recurrent neural netwokes with continuously distributed delays. Nonlinear Anal: Real World Appl. 7, 421–430 (2006)
23. Zhao, H., Wang, G.: Existence of periodic oscillatory solution of reaction-diffusion neural networks with delays. Phys. Lett. A 343, 372–383 (2005)
24. Lu, G.: Global exponential Stability and periodicity of reaction-diffusion delayed recurrent neural networks with Dirichlet Boundary conditons. Chaos Solitons & Fractals 35, 116–125 (2008)
25. Schiff, S., Jerger, K., Duong, D., et al.: Controlling chaos in the brain. Nature 370, 615–620 (1994)
26. Liao, X., Wong, K., Leung, C., et al.: Hopf bifurcation and chaos in a single delayed neuron equation with non monotonic activation function. Chaos Solitons & Fractals 12, 1535–1547 (2001)

Exponential Synchronization of Delayed Fuzzy Cohen-Grossberg Neural Networks with Reaction Diffusion Term

Teng Lv[1] and Ping Yan[2,⋆]

[1] Teaching and Research Section of Computer,
Artillery Academy, Hefei, 230031, P.R. China
[2] School of Science,
Anhui Agricultural University, Hefei, 230036, P.R. China
want2fly2002@163.com

Abstract. In this paper, the exponential synchronization of delayed fuzzy Cohen-Grossberg neural networks is discussed. Some sufficient conditions are given to ensure the exponential synchronization of the drive-response delayed fuzzy cellular neural networks.

Keywords: exponential synchronization, fuzzy Cohen-Grossberg neural networks, reaction diffusion term, Lyapunov functional.

1 Introduction

Fuzzy cellular neural networks (FCNN), introduced in 1996 by Yang et.al. [1], combines fuzzy logic with the traditional cellular neural networks (CNN). The dynamical behaviors of FCNN without delays or without diffusion terms have been discussed [1]-[3]. Cohen-Grossberg neural networks(CGNN) were first introduced by Cohen and Grossberg in [4]. The class of networks includes the traditional neural networks as special cases and has many good applications in parallel computation, associative memory and optimization problems, etc. Because these applications heavily depend on the dynamical behaviors of the networks and the analysis of the dynamical behaviors is the necessary step to design of neural networks, many results on the dynamical behaviors of the networks can be found in [5]-[6]. Just as FCNN, combining fuzzy logic with the traditional CGNN, we can obtain the fuzzy Cohen-Grossberg neural networks(FCGNN). The dynamical behaviors of FCGNN without diffusion term have been studied in [7]-[8].

Since Pecora and Carroll introduced chaos synchronization by proposing the drive-response concept in 1990s [9]-[10], the synchronization of coupled chaos systems and chaos neural networks has been received considerable attention [11]-[14].

⋆ Corresponding author.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 57–63, 2010.

However, As time delays may led to bifurcation, oscillation, divergence or instability which may be harmful to a system, thus, the study of neural dynamics with consideration of the delayed problem becomes extremely important to manufacture high quality neural networks. Therefore, the models with time-varying delays and continuous distributed delays are more appropriate to the synchronization of neural networks [15]-[16].

To the best of our knowledge, few authors have considered exponential synchronization of reaction-diffusion FCGNN with mixed delays. But it is important in theories and applications. In this paper, we will investigate exponential synchronization of a class of reaction-diffusion FCGNN with time-varying delays.

This paper is organized as follows. In section 2, model description and main result are given. In section 3, by employing Sobolev inequality techniques and constructing suitable Liapunov functional, some sufficient conditions are obtained to ensure the exponential synchronization of a class of reaction-diffusion FCGNN. Section 4 is the conclusions of the paper.

2 Model Description and Main Result

Consider the following delayed reaction-diffusion fuzzy neural networks with general boundary conditions:

$$\begin{aligned}\frac{\partial u_i(x,t)}{\partial t} &= \sum_{k=1}^{m}\frac{\partial^2 u_i(x,t)}{\partial x_k^2} - a_i(u_i(x,t))\cdot\Big[b_i(u_i(x,t)) \\ &\quad - \bigwedge_{j=1}^{n} k_{ij}g_j(u_j(x,t-\theta_{ij}(t))) \\ &\quad - \bigwedge_{j=1}^{n} r_{ij}\mu_j - \bigvee_{j=1}^{n} s_{ij}\mu_j - I_i\Big],\ t>0 \qquad (1)\\ u_i(x,t) &= \varphi_i(x,t), \quad -\tau < t \le 0,\ x\in\Omega,\\ u_i(x,t) &= 0, \quad (x\in\partial\Omega,\ t>0).\end{aligned}$$

Where $i = 1,\cdots,n$. Ω is a bounded open domain in $\mathbb{R}^m$ with smooth boundary $\partial\Omega$ and $mes\Omega > 0$ denotes the measure of Ω. n is the number of neurons in the networks. $u(x,t) = (u_1(x,t),\cdots,u_n(x,t))^T$, $u_i(x,t)$ denotes the state of the ith neural unit at time t and in space $x\in\Omega$. a_i and b_i represent an amplification function and an appropriately behaved function, respectively. k_{ij} is constant and represents element of fuzzy feedback MIN template with time-varying delays. r_{ij} and s_{ij} are constants and represent element of fuzzy feed-forward MIN template and element of fuzzy feed-forward MAX template, respectively. $\theta_{ij}(t)$ corresponds to the transmission delays along the axon of the jth neuron form the ith neuron at time t and satisfies

$$0 \le \theta_{ij}(t) \le \tau,\ \dot{\theta}_{ij}(t) < \rho < 1,$$

where $\tau > 0$ and $\rho < 1$ are given constants. $g_i(\cdot)$ is the activation function with $g_i(0) = 0,\ (i = 1,\cdots,n)$. μ_i and I_i denote input and bias of the ith neuron,

respectively. $\varphi(x,t) = (\varphi_1(x,t), \cdots, \varphi_n(x,t))^T$ (where $\varphi_i(x,t)$ is given smooth function defined on $\Omega \times (-\tau, 0]$) with the following norm

$$\|\varphi\|_2 = \sqrt{\sum_{i=1}^{n} \int_{\Omega} |\varphi_i(x,\cdot)|_\tau^2 dx},$$

where $|\varphi_i(x,\cdot)|_\tau = \sup_{-\tau<s\leq 0} |\varphi_i(x,s)|$.

Throughout the paper, we always assume that system (1) has a smooth solution $u(t,x)$ with the norm $\|u(\cdot,t)\|_2 = \sqrt{\sum_{i=1}^{n} \int_{\Omega} |u_i(x,t)|^2 dx}$, $\forall t \in [0, +\infty)$.

Let $v = (v_1, \cdots, v_n)^T$ be the control input vector and v_i stands for the external control input that will be appropriately designed for an certain control objective. In this paper, the control input vector $v = (v_1, \cdots, v_n)^T$ is assumed to take as the follows.

$$(v_1, \cdots, v_n)^T = M\Big(u_1(x,t) - \widetilde{u}_1(x,t), \cdots, u_n(x,t) - \widetilde{u}_n(x,t)\Big)^T,$$

where $M = (M_{ij})_{n\times n}$ is the controller gain matrix and will be appropriately chosen for synchronization in both drive system and response system. $\widetilde{u}_i(x,t), (i = 1, \cdots, n)$ satisfy the following response neural networks:

$$\begin{aligned}
\frac{\partial \widetilde{u}_i(x,t)}{\partial t} &= \sum_{k=1}^{m} \frac{\partial^2 \widetilde{u}_i(x,t)}{\partial x_k^2} - a_i(\widetilde{u}_i(x,t)) \cdot \Big[b_i(\widetilde{u}_i(x,t)) \\
&\quad - \bigwedge_{j=1}^{n} k_{ij} g_j(\widetilde{u}_j(x, t-\theta_{ij}(t))) \\
&\quad - \bigwedge_{j=1}^{n} r_{ij}\mu_j - \bigvee_{j=1}^{n} s_{ij}\mu_j - I_i\Big] + v_i(x,t), \ t > 0, \qquad (2) \\
\widetilde{u}_i(x,t) &= \psi_i(x,t), \quad -\tau < t \leq 0, \ x \in \Omega, \\
\widetilde{u}_i(x,t) &= 0, \quad (x \in \partial\Omega, \ t > 0).
\end{aligned}$$

Where $i = 1, \cdots, n$. $\psi = (\psi_1, \cdots, \psi_n)^T$, $\psi_i(x,s)(i = 1, \cdots, n)$ are bounded smooth functions defined on $\Omega \times (-\tau, 0]$.

In this paper, we always assume that:

$(H1)$ There exist positive constants m_i and M_i such that $0 < m_i \leq a_i(u_i) \leq M_i$, $i = 1, \cdots, n$. Moreover, for any $i = 1, \cdots, n$, $a_i(\cdot)$ is differentiable and there exists positive constant $\Gamma_i > 0$ such that $\Gamma_i = \sup_{x\in\mathbb{R}}\{\dot{a}_i(x)\} > 0$.

$(H2)$ $b_i(\cdot)$ is differentiable and there exists a positive $G_i > 0$ such that $0 \leq b_i \leq G_i$ for all $i = 1, \cdots, n$. Furthermore, $\dot{b}_i(\cdot)$ is the derivative of $b_i(\cdot)$ and $\alpha_i = \inf_{x\in\mathbb{R}}\{\dot{b}_i(x)\} > 0$.

$(H3)$ There exist positive constants $\beta_i > 0$ and $\Lambda_i > 0$ for any $i = 1, \cdots, n$ such that

$$\|f_i\| = \sup_u |f_i(u)| \leq \Lambda_i, \quad |g_i(x_i) - g_i(y_i)| \leq \beta_i |x_i - y_i|, \ \forall x_i, y_i \in \mathbb{R}.$$

$(H4)$ There exist constants $\gamma_{ij}, \eta_{ij}(i,j=1,\cdots,n)$ such that for all $i=1,\cdots,n$, it holds

$$-2\alpha_i m_i + \Gamma_i\Big[G_i + |I_i| + \sum_{j=1}^{n}\Big(|k_{ij}|\Lambda_j + |\gamma_{ij}\mu_j| + |\gamma_{ij}\mu_j|\Big)\Big] + \sum_{j=1}^{n}\Big[\Big(\beta_j|k_{ij}|\Big)^{2\gamma_{ij}}\Big] + \sum_{j=1}^{n}\Big[\frac{1}{1-\rho}\Big(\beta_i|k_{ji}|\Big)^{2-2\gamma_{ji}}\Big] + \sum_{j=1}^{n}\Big[|M_{ij}|^{2\eta_{ij}} + |M_{ji}|^{2-2\eta_{ji}}\Big] < 0.$$

Definition 1. The drive system (1) and the response system (2) are said to be exponentially synchronized if there exist positive constants $\gamma > 0$ and $\varepsilon > 0$ such that $\|u(\cdot,t) - \widetilde{u}(\cdot,t)\|_2 \le \gamma\|\varphi - \psi\|_2 e^{-\varepsilon t}, \forall t \ge 0$, where the constant ε is said to be the exponential synchronization rate.

The main result is the following theorem.

Theorem 1. *If* $(H1)-(H4)$ *hold, then the drive-response neural networks* (1) *and* (2) *are exponential synchronization.*

3 The Synchronization of the Drive and Response Neural Networks

Let us define the synchronization error signal $\varepsilon_i(x,t) = u_i(x,t) - \widetilde{u}_i(x,t)$, where $u_i(x,t)$ and $\widetilde{u}_i(x,t)$ are the ith state variable of the drive and response neural networks, respectively. Therefore, the dynamics error between (1) and (2) can be expressed as the follows.

$$\begin{aligned}
\frac{\partial\varepsilon_i(x,t)}{\partial t} &= \sum_{k=1}^{m}\frac{\partial^2\varepsilon_i(x,t)}{\partial x_k^2} - a_i(u_i(x,t))\cdot\Big[b_i(u_i(x,t)) \\
&\quad - \bigwedge_{j=1}^{n} k_{ij}g_j(u_j(x,t-\theta_{ij}(t))) - \bigwedge_{j=1}^{n} r_{ij}\mu_j - \bigvee_{j=1}^{n} s_{ij}\mu_j - I_i\Big] \\
&\quad + a_i(\widetilde{u}_i(x,t))\cdot\Big[b_i(\widetilde{u}_i(x,t)) - \bigwedge_{j=1}^{n} k_{ij}g_j(\widetilde{u}_j(x,t-\theta_{ij}(t))) \\
&\quad - \bigwedge_{j=1}^{n} r_{ij}\mu_j - \bigvee_{j=1}^{n} s_{ij}\mu_j - I_i\Big] - v_i(x,t), \\
\varepsilon_i(x,t) &= \varphi_i(x,t) - \psi_i(x,t), \quad -\tau < t \le 0, \\
\varepsilon_i(x,t) &= 0, \quad (x \in \partial\Omega).
\end{aligned}$$

Where $i=1,\cdots,n$ and the control input vector $v(x,t) = (v_1(x,t),\cdots,v_n(x,t))^T$ has been defined above.

Then we have

$$\frac{\partial\varepsilon_i(x,t)}{\partial t} = \sum_{k=1}^{m}\frac{\partial^2\varepsilon_i(x,t)}{\partial x_k^2} - \Big(a_i(u_i(x,t)) - a_i(\widetilde{u}_i(x,t))\Big)\cdot\Big[b_i(u_i(x,t))$$

$$- \bigwedge_{j=1}^{n} k_{ij} g_j(u_j(x,t-\theta_{ij}(t))) - \bigwedge_{j=1}^{n} r_{ij}\mu_j - \bigvee_{j=1}^{n} s_{ij}\mu_j - I_i\Big] \tag{3}$$

$$-a_i(\widetilde{u}_i(x,t)) \cdot \Big[b_i(u_i(x,t)) - b_i(\widetilde{u}_i(x,t))$$

$$- \bigwedge_{j=1}^{n} k_{ij}\Big(g_j(u_j(x,t-\theta_{ij}(t))) - g_j(\widetilde{u}_j(x,t-\theta_{ij}(t)))\Big)\Big] - v_i(x,t),$$

$$\varepsilon_i(x,t) = \varphi_i(x,t) - \psi_i(x,t), \quad -\tau < t \le 0,$$

$$\varepsilon_i(x,t) = 0, \quad (x \in \partial\Omega).$$

By $(H4)$, there exists a sufficiently small positive constant $\lambda < \min_i\{a_i m_i\}$ such that

$$W_i = \lambda - \alpha_i m_i + \frac{1}{2}\Big\{\Gamma_i\Big[G_i + |I_i| + \sum_{j=1}^{n}\Big(|k_{ij}|\Lambda_j + |\gamma_{ij}\mu_j| + |\gamma_{ij}\mu_j|\Big)\Big] \tag{4}$$

$$+ \sum_{j=1}^{n}\Big[\Big(\beta_j|k_{ij}|\Big)^{2\gamma_{ij}} + \frac{e^{2\lambda\tau}}{1-\rho}\Big(\beta_i|k_{ji}|\Big)^{2-2\gamma_{ji}} + |M_{ij}|^{2\eta_{ij}} + |M_{ji}|^{2-2\eta_{ji}}\Big]\Big\} \le 0.$$

Taking Liapunov funcional as follows:

$$V(t) = \sum_{i=1}^{n}\int_{\Omega}\Big[\varepsilon_i^2(x,t)e^{2\lambda t} + \sum_{j=1}^{n}\frac{\big(\beta_j|k_{ij}|\big)^{2-2\gamma_{ij}}}{1-\rho}\cdot\int_{t-\theta_{ij}(t)}^{t}\varepsilon_j^2(x,s)e^{2\lambda(s+\tau)}ds\Big]dx.$$

Calculating $D^+V(t)$ along system (3), we have

$$D^+V(t) \le 2e^{2\lambda t}\sum_{i=1}^{n} W_i\int_{\Omega}\varepsilon_i^2(x,t)dx \le 0,$$

where W_i is defined in (4). It means that

$$V(t) \le V(0) = \sum_{i=1}^{n}\int_{\Omega}\Big[\varepsilon_i^2(x,0) + \sum_{j=1}^{n}\frac{\big(\beta_j|k_{ij}|\big)^{2-2\gamma_{ij}}}{1-\rho}\int_{-\theta_{ij}(0)}^{0}\varepsilon_j^2(x,s)e^{2\lambda(s+\tau)}ds\Big]dx.$$

Let

$$\Upsilon = 1 + \frac{e^{2\lambda\tau}-1}{2\lambda(1-\rho)}\sum_{i,j=1}^{n}\big(\beta_j|k_{ij}|\big)^{2-2\gamma_{ij}} > 1.$$

For $V(t) \ge \|\varepsilon(\cdot,t)\|_2^2 e^{2\lambda t}$, it holds $\|\varepsilon(\cdot,t)\|_2^2 \le \Upsilon\|\varphi(\cdot,t)-\psi\|_2^2 e^{-2\lambda t}$, that is $\|u(\cdot,t)-\widetilde{u}(\cdot,t)\|_2^2 \le \Upsilon\|\varphi-\psi\|_2^2 e^{-2\lambda t}$, by Definition 1, we know that the drive-response neural networks can be exponentially synchronized. So we finish the proof of Theorem 1.

4 Conclusion

In this paper, the exponential synchronization of reaction diffusion fuzzy Cohen-Grossberg neural networks with time-varying delays is studied. Some sufficient conditions expressed by algebra inequalities have been given to ensure the exponential synchronization of reaction diffusion fuzzy Cohen-Grossberg neural networks. The methods used here can be used to deal with the exponential synchronization of general problems. The result obtained in this paper is still true for other delayed reaction-diffusion fuzzy neural networks.

Acknowledgment

The authors would like to thank Natural Science Foundation of Artillery Academy and College Science Research Plan Project of Xinjiang Uighur Autonomous Region(No.XJEDU2007I03 and No.XJEDU2007G01).

References

1. Yang, T., Yang, L.: The global stability of fuzzy cellular neural network. IEEE Trans. on Circuits and Systems-I: Fundamental Theory and Applications 43(10), 880–883 (1996)
2. Yang, T., Yang, L.: Fuzzy cellular neural network: a new paradigm for image processing. Int. J. Circuit Theory Appl. 25(6), 469–481 (1997)
3. Wang, S., Fu, D., Xu, M., Hu, D.: Advanced fuzzy cellular neural network: application to CT liver images. Artificial Intelligence in Medicine 39, 65–77 (2007)
4. Cohen, M., Grossberg, S.: Absolute stability and global pattern formation and parallel memory storage by competitive neural networks. IEEE Trans. Syst. Man Cybernet SMC-13, 15 (1983)
5. Chen, Z., Zhao, H., Ruan, J.: Dynamic analysis of high-order Cohen-Grossberg neural networks with time delay. Chaos Solitons Fractals 24, 421–430 (2006)
6. Jiang, H., Cao, J.: Dynamics of Cohen-Grossberg neural networks with time-varying delays. Phys. Lett. A 354, 414–422 (2006)
7. He, D., Xu, D.: Attracting and invariant sets of fuzzy Cohen-Grossberg neural networks with time-varying delays. Phys. Lett. A 372(47), 7057–7062 (2008)
8. Xiang, H., Cao, J.: Periodic Oscillation of Fuzzy Cohen-Grossberg Neural Networks with Distributed Delay and Variable Coefficients. J. Appl. Math. 2008, 18 (2008), doi:10.1155/2008/453627
9. Pecora, L., Carroll, T.: Synchronization in chaotic systems. Phys. Rev. Lett. 64(8), 821–824 (1990)
10. Pecora, L.M., Carroll, T.L., Johnson, G.A., Mar, D.J., Heagy, J.: Fundamentals of synchronization in chaotic systems, concepts and applications. Chaos 7(4), 520–543 (1997)
11. Sun, Y., Cao, J.: Adaptive lag synchronization of unknown chaotic delayed recurrent neural networks with noise perturbation. Phys. Lett. A 364(3-4), 277–285 (2007)
12. Xiong, W., Xie, W., Cao, J.: Adaptive exponential synchronization of delayed chaotic networks. Physica A 370(2), 832–842 (2006)

13. Cheng, C., Liao, T., Yan, J., Hwang, C.: Exponential synchronization of a class of neural networks with time-varying delays. IEEE Tran. Systems Man Cybernetics Part B: Cybernetics 36(1), 209–215 (2006)
14. Wang, Y., Cao, J.: Synchronization of a class of delayed neural networks with reaction-diffusion terms. Phys. Lett. A 369(3), 201–211 (2007)
15. Wang, J., Lu, J.: Globally exponential stability of fuzzy cellular neural networks with delays and reaction-diffusion terms. Chaos Solitons & Fractals 38, 878–885 (2008)
16. Huang, T.: Exponential stability of delayed fuzzy cellular neural networks with diffusion. Chaos Solitons & Fractals 31, 658–664 (2007)
17. Gilbarg, D., Trudinger, N.: Elliptic partial differential equations of second order, 3rd edn. Springer, Berlin (2001)

Magnetic Field Extrapolation Based on Improved Back Propagation Neural Network

Li-ting Lian, Chang-han Xiao, Sheng-dao Liu, Guo-hua Zhou, and Ming-ming Yang

School of Electrical and Information Engineering, Naval Univ. of Engineering, Wuhan, Hubei, 430033, China

Abstract. Magnetic anomaly created by ferromagnetic ships may make them vulnerable to detections and mines. In order to reduce the anomaly, it is important to evaluate magnetic field firstly. Underwater field can be measured easily, but upper air field is hard to be got. To achieve it, a model able to predict upper air magnetic field from underwater measurements is required. In this paper, a Back Propagation (BP) model has been built and it can escape from local optimum thanks to optimizing the initial weights and threshold values by Particle Swarm Optimization (PSO) algorithm. The method can avoid many problems from linear model and its high accuracy and good robustness have been tested by a mockup experiment.

Keywords: BP, magnetic field extrapolation, PSO, upper air field, underwater field.

1 Introduction

Magnetic silencing techniques are vital for ships especially for submarines. The magnetic anomaly created by ferromagnetic submarines may compromise their invisibility. Many antisubmarine planes find objects by sonar detections and locate them exactly by magnetism detections. Therefore, for decades, marines worldwide are looking for reducing magnetic anomaly [1]. Before achieving it, it is necessary for the ship to evaluate its magnetic anomaly. Underwater field can be measured easily, but upper air field is hard to be got. To achieve it, a model able to predict upper air field from underwater measurements is required. Many researchers settle this kind of problem by numerical approaches like finite element method, boundary element method, integral equation method and method of virtual source [2~5]. With the developments of neural network, some researchers study the magnetic problem by neural network. A. Salem [6] has ascertained the positions of Unexploded Ordnance (UXO) thanks to a neural network and G.H. Zhou [7] has studied the localization of a magnetic object based on Hopfield neural network.

In this paper, we proposed a Back Propagation (BP) neural network model to solve it. The model can escape local optimum thanks to optimizing the initial weight and threshold values by Particle Swarm Optimization (PSO) algorithm. The method can

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 64–70, 2010.

avoid many problems from linear model and its high accuracy and good robustness has been tested by a mockup experiment.

2 The Traditional Method of Magnetic Field Extrapolation

In this section, we are interested in the calculation of magnetic anomaly evaluation with the method of virtual source. The magnetic field is generated by magnetic sources. When the observation points are far away enough from magnetic sources, the sources can be equivalent to magnetic dipoles. For the same ship, the upper air field and underwater field are generated by the same magnetic sources. The difference is the observation points' location (see Fig.1).

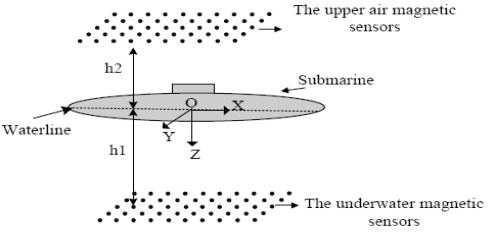


Fig. 1. The sketch map of submarine and sensors' location

The magnetic flux density $\boldsymbol{B}$ in air region generated by a single magnetic dipole can be expressed in terms of magnetic moments $\boldsymbol{m}$:

$$\boldsymbol{B}=\frac{\mu_0}{4\pi}\left(\frac{3(\boldsymbol{m}\cdot\boldsymbol{r})\boldsymbol{r}}{r^5}-\frac{\boldsymbol{m}}{r^3}\right) \tag{1}$$

where $\boldsymbol{r}$ is the vector distance directed from the magnetic dipoles to the observation points, μ_0 is the magnetic permeability of free space. Take the point P located in the underwater measurement plane (see Fig.2) as an example, carrying out simple decompose, we obtain the equation of observation point P:

$$\begin{cases} \boldsymbol{B}_{xp}=\dfrac{\mu_0}{4\pi r_p^5}[(3(x_p-x_0)^2-r_p^2)\boldsymbol{m}_x+3(x_p-x_0)(y_p-y_0)\boldsymbol{m}_y+3(x_p-x_0)(z_p-z_0)\boldsymbol{m}_z] \\ \boldsymbol{B}_{yp}=\dfrac{\mu_0}{4\pi r_p^5}[3(x_p-x_0)(y_p-y_0)\boldsymbol{m}_x+(3(y_p-y_0)^2-r_p^2)\boldsymbol{m}_y+3(y_p-y_0)(z_p-z_0)\boldsymbol{m}_z] \\ \boldsymbol{B}_{zp}=\dfrac{\mu_0}{4\pi r_p^5}[3(x_p-x_0)(z_p-z_0)\boldsymbol{m}_x+3(\mathrm{y}_p-\mathrm{y}_0)(z_p-z_0)\boldsymbol{m}_y+(3(z_p-z_0)^2-r_p^2)\boldsymbol{m}_z] \end{cases} \tag{2}$$

Writing at every source, this equation leads to a linear system whose solution gives the magnetic moment vectors M of each source. The magnetic matrix is called A, and the vector of magnetic flux density is called C:

$$A \cdot M = C \tag{3}$$

The equation is also suitable for the points in the upper air measurement plane. Previous researchers start from the measurement of underwater field, then determine the moment M of the sources by inverse problem, and last the upper air magnetic field will be calculated directly by forward modeling. The precision in calculation can is often low except adding other information such as the measure- ments on enveloping surface. All said above has proposed an indirect method of magnetic anomaly evaluation, a neural network method that finds the relationship between upper air and underwater field directly will be introduced in next section.

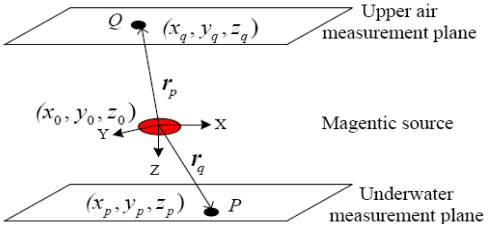


Fig. 2. The sketch map of observation points and magnetic source

3 The Model of Magnetic Field Extrapolation Based on Improved BP Neural Network

3.1 The Classical BP Neural Network

Recently, artificial neural networks have been found wide applications in many areas to deal with some nonlinear problem thanks to its strong ability to work as biological neural system. There are many effective models such as Hopfield neural network, RBF neural network, and BP neural network which is the most practicable network thanks to its simple structure and strong learning ability. It is said that three layers feed forward network can approximate any nonlinear function if the activation function of hidden layer node is Sigmoid [8]. The magnetic field extrapolation can be reduced to a nonlinear problem, so the BP network model which has a single hidden layer has been built to settle the problem. The structure of the model is shown as Fig.3. Where the input vector is underwater magnetic field vector; the output vector is upper air magnetic field vector; R is the dimension of input; S1 is the number of hidden layer nerve cells; S2 is the dimension of output; the transfer function of hidden layer is tansig; the transfer function of output layer is purelin. But classical BP algorithm is apt to get a local optimum and sensitive to initial weight value and bias value. In order to improve its robustness, a hybrid optimal algorithm PSO-BP is employed.

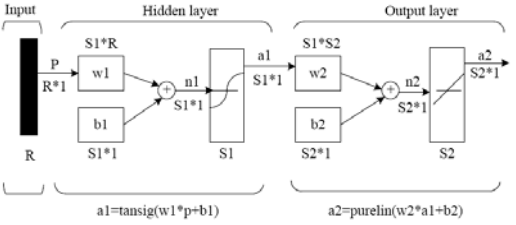


Fig. 3. A classical BP neural network

3.2 Improved BP Network Based on PSO Algorithm

The reason for bad robustness is that the network starts from different initial weight values and threshold values. We can get optimal initial weight values and threshold values by the random search ability of PSO algorithm [9-10].

PSO was proposed by Dr Eberhart and Kennedy in 1995[11]. Firstly, initialize a team of random particles and the velocity and then search the optimal solution of particles by iteration; every particle updates velocity and direction by P_{best} (the single extreme found by each particle) and G_{best} (the global extreme found by the whole particle swarm) in every iteration, the regulating rule of the position and the velocity follow (4) and (5):

$$V(t+1) = wV(t) + C_1 \times R_1 \times [P_{best}(t) - X(t)] + C_2 \times R_2 \times [G_{best}(t) - X(t)] \tag{4}$$

$$X(t+1) = X(t) + V(t+1) \tag{5}$$

where V is the velocity of different iterations, X is the position of the particle, C_1 and C_2 called accelerated coefficients are both positive value, w is the inertia mass that controls the impact of the previous velocity of the particle, R_1 and R_2 are two random numbers with the range of [0, 1].

The procedure of PSO-BP is listed below:

step1: construct the BP network's basic structure, initialize the PSO parameters, and generate initial particles;

step2: map particles as the initial weight value and threshold value of network and then train the net to get root--mean-square error of the net;

step3: follow the rule of PSO to regulate the velocity and direction to find the optimal weight value and threshold value which make the root-mean-square error of net is least;

step4: train the network via the optimal weight value and threshold value till the network has been steady, then use the testing samples to valid the network's forecast accuracy and its generalization ability.

4 Experiments to Validate the Nonlinear Model

4.1 Experimental Design

We are interested in the model able to predict upper air field from underwater ones. Five triaxial sensors have been placed 36.8 centimeters under a mockup to get net's

input vectors and five triaxial sensors have been placed 40.8 centimeters up the mockup to get output vectors (see Fig.4). If we move the mockup and keep sensors immobile, there are five measuring lines (from line (1) to line (5)) in both planes (see Fig. 4) and 41 measuring points in every line. Measuring interval is 10 centimeter, there are 41×5 measuring points(the order is 1~41 form left to right)for each plane.

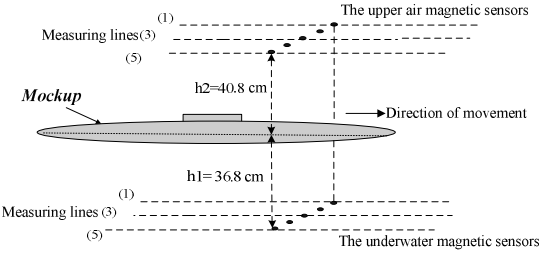


Fig. 4. The sketch map of sensors' position arrangement

4.2 Fundamental Parameters of BP Neural Network

Separate measurements of two measuring planes into part A (25×5) and part B (16×5). The rule is that we choose the points (the even with the range of [4, 34]) of two planes as part B, the other points as part A.

Input and output vectors of training: Part A of two planes is selected as input and output samples respectively;

Input of testing: Part B of underwater plane is selected as input of testing sample,

Theoretic output: B of upper air field is selected as the theoretic output vectors (measured field).

The node of hidden layer: The number is determined by $S1=\sqrt{R\times S2}+k$ where R=5, S2=5(see fig.3), k is a natural number. The preferable number (S1) is equal to 6 after trying many times. The training algorithm is Levenberg Marquardt (LM).

4.3 Fundamental Parameters of PSO Algorithm

The population size is 30, the dimension of particle is the total number of weight values and threshold values, and we choose net's root-mean-square error as adaptation function. The Other parameters areC_1=2, C_2=2, w=0.8.

4.4 Train and Evaluate the Network

After setting parameters, train the network according to PSO-BP algorithm, and then input the vectors of testing samples, the field obtained by network then compared to measures field (after data normalization). The relatively root-mean -square error is selected as the criterion (Er=Rms(k)/max(t(k),k=1,2…5)), where Rms (k) is the root-mean-square error of a measuring line and t(k) is the measurements outside. The network's high accuracy has been shown as Fig.5. The relatively root-mean-square error(Er) is less than 5%, which is better than the result that is less than 10% got by traditional methods.

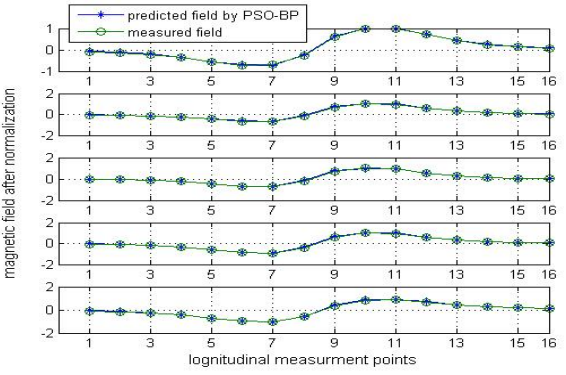


Fig. 5. Comparison of field predicted by PSO-BP and measured field of five measuring lines

5 Conclusion

We have developed the application of BP neural network in ship's magnetic field extrapolation and tested it by a mockup experiment. It presents advantage in comparison to traditional linear methods. First, it can avoid errors made by complicated ship's magnetic construction, mesh and inverse calculus from traditional linear methods. Second, by finding nonlinear relationship between underwater field and upper air field directly, we improve the accuracy of model and get good robustness thanks to optimizing the initial weight value and threshold value by PSO algorithm. A future work will deal with real ship's magnetic field extrapolation to verify the practicable application.

References

1. Vuillermet, Y., Chadebec, O., Coulomb, J.L., Rouve, L.-L., Cauffet, G., Bongiraud, J.P., Demilier, L.: Scalar Potential Formulation and Inverse Problem Applied to Thin Magnetic Sheets. J. IEEE Transaction on Magnetic 44, 1054–1058 (2008)
2. Chadebec, O., Coulomb, J.L., Bongirand, J.P., Cauffet, G., Thiec, P.L.: Recent Improvements for Solving Inverse Magnetostatic Problem Applied to Shin Shells. J. IEEE Transaction on Magnetic 38, 1005–1008 (2002)
3. Rakotoariso, H.L., Ardon, V., Chadebec, O., Delinchant, B., Guerin, S., Coulomb, J.L.: Formal Sensitivity Computation of Magnetic Moment Method. J. IEEE Transaction on Magnetic 44, 1014–1017 (2008)
4. Guibert, A., Chadebec, O., Coulomb, J.-L., Rannou, C.: Ships Hull Corrosion Diagnosis from Close Measurements of Electric Potential in The Water. J. IEEE Transaction on Magnetic 45, 1828–1831 (2009)
5. Guo, C.B., Xiao, C.H., Liu, D.M.: Research on The Continuations of Magnetic Object Based on Integral Equation Method and Singular Value Decomposition. J. Acta Physica Sinica 57, 4182–4188 (2008) (in Chinese)

6. Salem, A., Ushijima, K.: Automatic Detection of UXO From Airborne Magnetic Data Using A Neural Network. J. Subsurface Sensing Technologies and Applications 2, 191–213 (2001)
7. Zhou, G.H., Xiao, C.H., Liu, S.D.: Magnetic Moments Estimation and Localization of A Magnetic Object Based on Hopfield Neural Network. J. Electromagnetic Field Problems and Application 21, 19–24 (2009)
8. Qiao, W.D.: Transformer Fault Diagnosis Based on SAPSO-BP Hybrid Algorithm. J. High Voltage Apparatus 44, 208–211 (2008) (in Chinese)
9. Wang, P., Huang, Z.Y., Zhang, M.G., Zhao, X.W.: Mechanical Property Prediction of Strip Model Based on PSO-BP Neural Network. J. Journal of Iron and Steel Research 15, 87–91 (2008)
10. Pan, F., Chen, J., Tu, X.Y., Fu, J.W.: Multilayered Feed Forward Neural Network Based on Particle Swarm Optimizer Algorithm. J. Journal of System Engineering and Electronics 16, 682–686 (2005)
11. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceeding of the IEEE International Conference on Neural Network, pp. 39–44. IEEE Press, Perth (1995)

Sparse Deep Belief Net for Handwritten Digits Classification

Jiongyun Xie[1], Hongtao Lu[2], Deng Nan[3], and Cai Nengbin[4]

[1,2] MOE-MS Key Laboratory for Intelligent Computing and Intelligent Systems
Dept. of Computer Science, Shanghai Jiaotong University, Shanghai, China
andingxie@gmail.com, lu-ht@cs.sjtu.edu.cn
[3,4] Shanghai Forensic Center
{dengnan1985,cainengbin}@sina.cn

Abstract. It has been shown that the Deep Belief Network is good at modeling input distribution, and can be trained efficiently by the greedy layer-wise unsupervised learning. Hoglak Lee et al. (2008) introduced a sparse variant of the Deep Belief Network, which applied the Gaussian linear units to model the input data with a sparsity constraint. However, it takes much more weight updates to train the RBM (Restricted Boltzmann Machine) with Gaussian visible units, and the reconstruction error is much larger than training an RBM with binary visible units. Here, we propose another version of Sparse Deep Belief Net which applies the differentiable sparse coding method to train the first level of the deep network, and then train the higher layers with RBM .This hybrid model, combining the advantage of the Deep architecture and the sparse coding model, leads to state-of-the-art performance on the classification of handwritten digits.

Keywords: Deep Belief Network, Restricted Boltzmann Machine, Sparse Coding.

1 Introduction

Restricted Boltzmann Machine[5] is a powerful generative model to model the data distribution.In earlier training method,prolong Gibbs sampling method was proposed to obtain the equilibrium distribution of the network.However, this was very time-consuming and samples from the equilibrium distribution generally have high variance since they come from all over the models distribution.Hinton et al.[5] propose a contrastive divergence method to train the RBM.Instead of running the Gibbs sampler to reach its equilibrium, their method needs only one Gibbs sampling iteration to approximate the data reconstruction.Although it is not exactly following the gradient of the log probability of the training data,contrastive-divergence has been shown to produce only a small bias for a large speed-up in training time,and the variance coming from the reconstruction is very small because the one-step reconstruction is very similar to the data.

The contrastive-divergence learning algorithm makes the RBM available to practical applications. Hinton et al.[4] proposed a greedy layer-wise training

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 71–78, 2010.

method to train the deep belief net.RBM was used here to train every layer of the deep architecture to provide the initial value for the whole network,then the deep network can be fine-tuned with supervised back-propagation.It performs excellent on the handwritten digits classification.

Hoglak Lee et al.[6] proposed a sparse RBM which applies the Gaussian(linear) visible units to model the input data and adds a regularization term that penalizes a derivation of the expected activation of the hidden units from a (low) fixed level p to produce sparse codes.As long as the p is sufficient small,the sparse RBM will learn the sparse (stroke like) features from modeling the input data as other sparse coding model. To train the sparse RBM,one needs only to apply the contrastive divergence update rule,followed by one step of gradient descent using the gradient of the regularization term.

Nonetheless,learning an RBM with Gaussian visible units can be slow, as it may require a much grater number of weight updates than an equivalent RBM with binary visible units. Moreover,the reconstruction error in training the RBM with Gaussian visible units is much large compared with the equivalent RBM with binary visible units.So we conjecture that contrastive-divergence may not be accurate enough to train the RBM with Gaussian units.

In this paper, we develop a different version of the sparse deep belief net.We train the first layer of deep belief net with the differentiable sparse coding scheme[13] instead of RBM.Then we take the sparse codes as the input to the higher layer of the deep belief network.Because the sparse codes produced by the sparse coding scheme are quasi binary,we can train the higher layer of the deep belief net with the standard binary RBM perfectly.

After the pre-training strategy,we obtained a sparse Deep Belief Net which keeps the powerful information of the sparse features from modeling the input data.We will show that this sparse Deep Belief Net perform very well on the discriminative task.

2 Restricted Boltzman Machine

Restricted Boltzmann Machine (RBM) is an undirected graphical model (Markov random field).It is a powerful generative model with one visible layer, one hidden layer and no intra-layer connections.

The energy function of RBM is

$$E(v,h) = -\sum_i v_i\theta_i - \sum_j h_j\lambda_j - \sum_{ij} W_{ij}v_ih_j \tag{1}$$

where the v and h represent the corresponding visible units and the hidden units.W is the matrix of the pair-wise weights.The corresponding joint distribution for the RBM is

$$p(v,h) = \frac{exp(-E(v,h))}{E_I} \tag{2}$$

where

$$E_I = \sum_{u,g} exp(-E(u,g)) \tag{3}$$

is the partition function of the model. From (2) we can compute

$$p(v) = \frac{1}{E_I} \cdot \sum_h (exp(-E(v,h))) \tag{4}$$

Given a set of N training cases $v_1, v_2, v_3 ... v_n$, we then train the parameter by maximizing

$$L = \sum_{i=1}^{n} \log P(v_i) \tag{5}$$

so as to minimize the energy of states drawn from the data distribution and raise the energy of states that are improbable given the data. To calculate the gradient of L with respect to the parameter θ, we obtain

$$\frac{\partial L}{\partial \theta} = -\langle \frac{\partial E}{\partial \theta} \rangle_{p_0(h,v)} + \langle \frac{\partial E}{\partial \theta} \rangle_{p_\infty(h,v)} \tag{6}$$

where p_∞ represents the equilibrium distribution of the model,which can be obtained by running prolong Gibbs sampling, p_0 represents the distribution of the model when the visible units are clamped on the training data.However, to get the equilibrium distribution is very time-consuming and would produce samples with high variance.Hinton et.al[5] introduced the contrastive divergence which has been found to be efficient in training the energy-based model.The idea is that instead of running the Gibbs sampler to its equilibrium distribution, they get the samples by running only one (or a few) Gibbs sampling.

2.1 Sparse RBM with Gaussian Visible Units

Welling et.al[9] introduced Gaussian hidden units which were used in an information retrieval task. Here we are interested in applying the linear units to handle with the continuous input. The energy function for an RBM with Gaussian visible unit is

$$E(v,h) = \sum_{i \in pixels} \frac{(v_i - b_i)^2}{2{\sigma_i}^2} - \sum_{j \in features} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j u_{ij} \tag{7}$$

where σ_i is the standard deviation of the Gaussian noise for unit i. The update rule is similar to the binary case.When the linear units have Gaussian noise with unit variance,the stochastic undate rule for the hidden unit remain the same,while the visible units i is to sample from a Gaussian with mean $b_i + \sum_j h_j W_{ij}$ and unit variance.When the variance is not 1,it need only minor adjustment which can be found in [4].

Hoglak Lee et.al[6] proposed a sparse version of the RBM which apply the Gaussian(linear) units to model the input data and add a regularization term to penalize the derivation of the expected activation of the hidden units from a(low) fixed level p, as follows:

$$\max\{\log P(v) - \gamma(p - E(h|v))\} \tag{8}$$

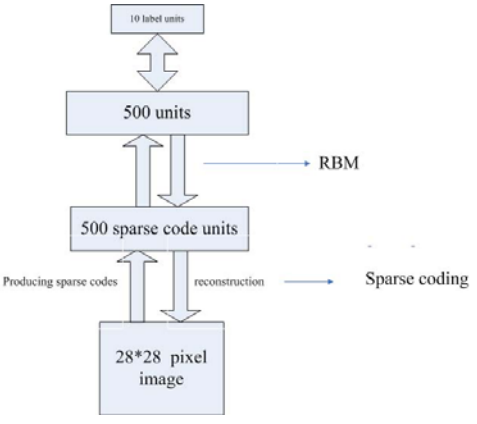


Fig. 1. The framework of our sparse DBN model. The first layer of the net is initialized with the sparse coding algorithm,and the higher layer is trained with RBM.

where v represents the training data,$E(h|v)$ is the expect activation of the hidden units.The training algorithm needs only small modification which can be found in [6].

However, the contrastive divergence learning may not be powerful enough to train RBM with Gaussian linear visible units.Compared with the RBM with binary units,the reconstruction error is much larger. We analyze this as follows, when trained with the binary units,the RBM only needs to model the input distribution that scaled in (0,1). However, when trained with the continuous valued inputs,the RBM has to model the unconstrained valued data distribution,and the energy surface is much more sophisticate,so it is much more difficult to model the manifold of and around the training samples.So we conjecture that the RBM with Gaussian linear visible units trained with contrastive divergence may not be sufficient accurate to be applied in a discriminative task.In the following section we propose an sparse coding strategy to train the first linear layer of the deep nets.

3 Sparse Deep Belief Net

Motivated by the sparse RBM,we propose an differentiable sparse-coding scheme that was used in [13] to train the first layer of the deep net.Then we take the sparse codes as the input of the higher layer,and train the weights of the higher layer with the standard RBM(see Figure 1).

3.1 Differentiable Sparse Coding

In sparse methods,the codes is forced to have only a few non-zero units while most units are zeros or close to zero most of the time.It has been shown that sparse over-complete representation have many theoretical and practical advantages,as is shown in [2] [7].In particular,sparse codes have good robustness to

noise and perform well in the classification tasks.Sparse representation is motivated by the organization of the cortex.The receptive fields in visual area V1 have been reasonably well described physiologically and can be characterized as being localized,oriented and bandpass which is similar to the filters learned from sparse coding.

The differentiable sparse-coding scheme[13] used here is to minimize the the following equation,with respect to the sparse codes W,

$$Loss = \sum_i \{(x_i - BW_i)^2 + \gamma D_p(W_i||p)\} \tag{9}$$

where x represents the input data and B is the basis of the model.$D_p(w||p)$ is the regularization function which measure the distance between the sparse code w and a parameter vector p,here we use the unnormalized KL-divergence[13]:

$$D_p(w||p) = \sum_j (w_j \log \frac{w_j}{p_j} - w_j + p_j) \tag{10}$$

When the constant vector p is sufficient close to zero,KL-divergence prior can approximate the L_1 prior[13]. L_1 prior has been shown to be good at producing sparse codes.However,L_1 prior does not produce differentiable MAP(Maximum-A-Posteriori) estimates.KL-divergence prior,which preserves the sparsity benefits of the L_1-regularization,is smooth that can produce much stable latent codes which lead to better classification performance. Additionally,because of the smoothness of the KL-divergence prior,the basis B,can be optimized discriminatively through gradient descent back-propagation[13].

3.2 Combining Deep Belief Net and Sparse Coding

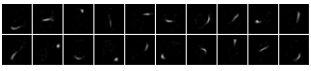
After training the first layer of the deep belief net using differentiable sparse coding,it can be easily integrated in the Deep Belief Net.We take the sparse codes as the input of the higher layer,then train the weights of the higher layer with the standard RBM.We can model the codes with the standard RBM perfectly because the codes we obtain are sparse and quasi binary.

To apply the Deep Belief Net in a classification task,unsupervised training is not enough.After training the hybrid model described above to initialize the weights of the deep belief net,we fine-tune the whole network with back-propagation. Because of the smooth KL-divergence prior,the gradient of the sparse codes w with respect to the basis B can be computed with the implicit differentiation,which can be found in[13].

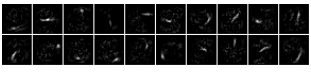
4 Experiments

4.1 Learning the Sparse Feature from Handwritten Digits

We applied the differentiable sparse coding scheme to the MNIST handwritten digit data set.Here we separated the original training set into training and



(a) Parts of the "strokes" like filters learned by sparse coding before back-propagation.



(b) The corresponding filters of the sparse-coding layer of the DBN after back-propagation.

Fig. 2. From the above two graph,we can see that even after the backpropagate,the first layer of the Deep Belief Network keeps most of the information learned from the sparse coding algorithm.

Table 1. Comparison of the classification performances on the MNIST database.Our sparse version of deep belief net(sparse coding+RBM) achieves the best result.SVM and multilayer neural network results are taken from http://yann.lecun.com/exdb/mnist/.On this data set,differences of 0.2% in classification error is statistically significant.

Model	Error
RBM+RBM	1.35%
Sparse RBM+RBM	1.42%
Sparse Coding+RBM	1.33%
Multilayer neural network	1.51%
SVM	1.40%

validation sets of 50000 and 10000 examples and used the standard test set of 10000 examples.We learned a sparse layer with 784 visible units and 500 hidden units.The learned bases are shown in Figure 2(a).In Figure 2(a) we can see that the basis found by the model roughly represent different strokes of which handwritten digits are comprised.This is consistent with results obtained by applying different sparse coding algorithms to learn sparse representations of this handwritten data set.

4.2 Learning the Sparse Deep Belief Net

By applying the 500 sparse codes obtained from the above algorithm as the input to the higher layer,we used the RBM model to train the higher layer of the net with 500 hidden units.The layer-by-layer pre-training provided us with the initial value of the whole network.At the fine-tuning stage,we added an extra layer with 10 units for the digit classification and proposed the conjugate gradient back-propagation to minimize the cross-entropy error function.Then we

obtained a 784-500-500-10 multilayer network.The experiment results are given in table 1.The RBM+RBM uses the standard RBM to initialize the whole Deep Belief network.Sparse RBM+RBM uses the sparse RBM to train the first layer of the Deep Belief Network,and sparse coding+RBM is the model we propose in this paper. We give as a comparison the results of a Gaussian kernel SVM and a regular neural network without pre-training.

We observe that the Deep Belief Networks with pretraining performs better than the regular neural network(random initialization),and our Sparse Deep Belief Network outperforms other DBN models.

Pre-training helps generalization because it ensures that most of the information in the initial weights come from modeling the handwritten digits.In our experiment,the filters learned in the sparse layer before and after fine-tuning are shown in Figure 2,from which one can see that back-propagation hasnt destroyed the sparse(stroke like) features. So even after the back-propagation,the network still keeps most of the information from modeling the input images.This explains why the Deep Belief Net model with pre-training perform better than the neural network without pre-training.

Our sparse version of Deep Belief Net achieves the best result on the MNIST handwritten data set,which implies the advantage of the sparse features on classification.It has been shown in many papers that sparse coding perform excellent on both the generative and discriminative models[12] [13].The sparse codes are much more robust than the original input. The sparse features represent the deep structure of the input data.

5 Conclusion

This paper is motivated by the need to develop good training algorithms for deep architectures.It has been show that layer-by-layer training would improve the generative power of the deep architectures as long as the number of the hidden units are sufficient large[1].Deep Belief Networks and the sparse coding models have many characters in common.They both exploit the deep structure of the input data.In this paper,we combine the advantage of Deep Belief Networks and Sparse coding together and construct a discriminative model with quiet good result on the classification of the handwritten digit data set.

For future work,we would like to investigate the model in various task,including the facial feature extraction and classification.We are also interested in applying the hybrid model in the generative task[13] [14].Deep Belief Networks have been shown to have powerful generative ability to model the distribution of the input data.If we combine it with the superior sparse features,maybe we will make fundamental improvement.

Acknowledgments. We thank Heng Luo for helpful discussions on Deep Belief Networks.Thanks to Huiyan Tu for discussions on the Sparse Coding. This work is supported by Shanghai Committee of Science and Technology (No.08JG05002) and 863 (No. 2008AA02Z310).

References

1. Hinton, G.E., Osindero, S., Yee-Whye, T.: A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18, 1527–1544 (2006)
2. Olshausen, B.A.: Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural images. Nature 381, 607–609 (1996)
3. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy Layer-Wise Training of Deep Networks. In: NIPS (2007)
4. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
5. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation 14, 1771–1800 (2002)
6. Lee, H., Ekanadham, C., Ng, A.Y.: Sparse Deep Belief Net Model for Visual Area V2. In: NIPS (2008)
7. Ranzato, M., Poultney, C., Chopra, S., LeCun, Y.: Efficient Learning of Sparse Representations with an Energy-based Model. In: NIPS (2006)
8. Ranzato, M., Boureau, Y.-L., LeCun, Y.: Sparse Feature Learning for Deep Belief Networks. In: NIPS (2007)
9. Welling, M., Rosen Zvi, M., Hinton, G.E.: Exponential Family Harmoniums with an Application to Information Retrieval. In: NIPS (2005)
10. Teh, Y.W., Welling, M., Osindero, S., Hinton, G.E.: Energy-based Models for Sparse Overcomplete Representations. Journal of Machine Learning Research 4, 1235–1260 (2003)
11. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient Sparse Coding Algorithm. In: NIPS (2006)
12. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught Learning: Transfer Learning from Unlabeled Data. In: ICML (2007)
13. Bradley, D.M., Bagnell, D.: Differentiable Sparse Coding. In: NIPS (2008)
14. Larochelle, H., Bengio, Y.: Classification using Discriminative Restricted Boltzmann Machine. In: ICML (2008)
15. Nair, V., Hinton, G.: Implicit Mixtures of Restricted Boltzmann Machine. In: NIPS (2008)
16. The MNIST database of handwritten digits, http://yann.lecun.com/exdb/mnist/

Multisensor Image Fusion Using a Pulse Coupled Neural Network

Yi Zheng[1] and Ping Zheng[2]

[1] School of Information and Electronic Engineering, Shandong Institute of Business and Technology, 264005 Yantai, China
zhengyi@sdibt.edu.cn
[2] Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, 430074 Wuhan, China
pzhhust@sohu.com

Abstract. Multisensor image fusion has its effective utilization for surveillance. In this paper, we utilize a pulse coupled neural network method to merge images from different sensors, in order to enhance visualization for surveillance. On the basis of standard mathematical model of pulse coupled neural network, a novel step function is adopted to generate pulses. Subjective and objective image fusion performance measures are introduced to assess the performance of image fusion schemes. Experimental results show that the image fusion method using pulse coupled neural network is effective to merge images from different sensors.

Keywords: image fusion, pulse coupled neural network, step function, performance assessment.

1 Introduction

Surveillance systems are currently being used for security surveillance, navigation as well as battlefield monitoring. The purpose of image fusion is to extract and synthesize information from multiple images in order to produce a more accurate, complete and reliable composite image of the same scene or target, so that the fused image is more suitable for human or machine interpretation. In this way, surveillance systems utilize the synergism of different unmated imaging sensors, such as a CCD camera and an infrared camera. In this paper, we consider a more restricted case of the multisensor fusion problem, i.e. we explore only the case of merging registered image pairs. And image fusion algorithms discussed in this paper are pixel-level.

A pulse coupled neural network (abbr. PCNN) is a biological inspired neural network developed by Eckhorn in 1990 and based on experimental observations of synchronous pulse bursts in cat visual cortex[1]. It is characterized by the global coupling and pulse synchronization of neurons. PCNN is different from what we generally mean by artificial neural networks in the sense that it does not train. PCNN is quite powerful and there are extensive image processing applications, such as segmentation, edge extraction, noise suppression, motion processing, and so on[2].

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 79–87, 2010.

PCNN has a history of some two decades but has solid applications in the field of image fusion only in the past ten years[3-9]. And it is still in underdevelopment stage and shows tremendous promise. In this paper, we explore characters of PCNN and utilize the PCNN model to perform image fusion in terms of gray level values.

The paper is organized as follows. In Section 2, we will illustrate the structure of PCNN in image fusion, enumerate the mathematical model of PCNN, and modify its step function of the mathematical model. The experiments are presented in Section 3. In the last section, some concluding remarks and future directions for study are given.

2 Image Fusion Based on PCNN

2.1 Structure of PCNN in Image Fusion

In a pulse coupled neural network, each neuron consists of three parts: the receptive field, the modulation field and the pulse generator. The simplified model of a PCNN neuron is shown in Fig. 1[10].

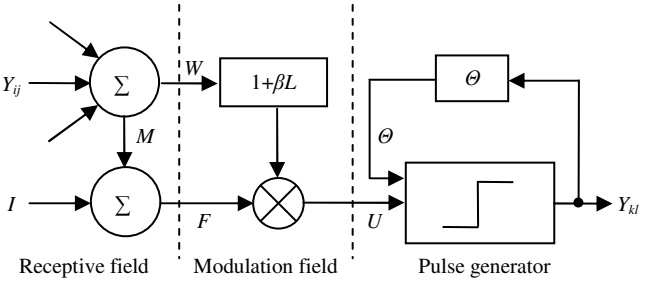


Fig. 1. Simplified model of a PCNN neuron

In Fig. 1, I, U, L, β is the external stimulus, the interior action, the linking input and the linking strength, respectively. F is the feeding part, which includes the input from both exterior and other neurons. PCNN is a feeding back network. Y_{ij} is the input from other neurons, and Y_{kl} is the output to other neurons. Θ is the variable threshold function. M and W are the matrix of linking weights of the feedback path and link path respectively.

Fig. 2 shows the connection model of a single neuron within the PCNN. PCNN is a single layer of pulse coupled neurons, and each neuron is corresponding to an input image pixel and an output image pixel.

When a pulse coupled neural network is applied to two dimensional image fusion, the image matrix corresponds to the neural network composed of $P \times Q$ PCNN neurons. And the gray level value of each pixel corresponds to the input of each neuron I. Each PCNN neuron receives inputs from its own stimulus and also from neighboring sources, i.e. outputs of other PCNN neurons in feeding radius is added to the input.

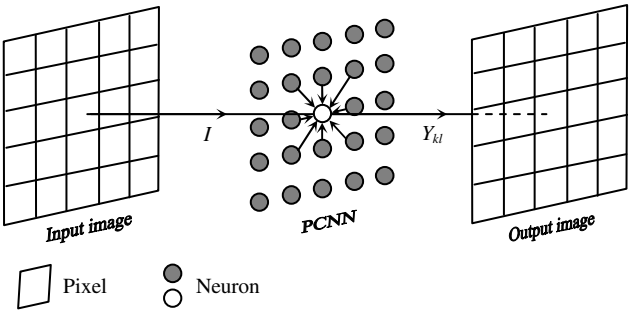


Fig. 2. Connection model of a PCNN neuron

2.2 Mathematical Model of PCNN

The multi-spectral PCNN is a set of parallel PCNNs each operating on a separate channel of the input with both inter- and intra-channel linking. In the surveillance system, input image pairs are visible and infrared images. Thus the number of channel is 2. Discrete equations of each neuron are the following.

$$F_{ij}(n) = e^{-\alpha_F} F_{ij}(n-1) + V_F \sum_{kl} M_{ijkl} Y_{kl}(n-1) + I_{ij} \tag{1}$$

$$L_{ij}(n) = e^{-\alpha_L} L_{ij}(n-1) + V_L \sum_{kl} W_{ijkl} Y_{kl}(n-1) \tag{2}$$

$$U_{ij}(n) = F_{ij}(n)\left(1 + \beta L_{ij}(n)\right) \tag{3}$$

$$Y_{ij}(n) = \begin{cases} 1, & \text{if } U_{ij}(n) > \Theta_{ij}(n-1) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\Theta_{ij}(n) = e^{-\alpha_\Theta} \Theta_{ij}(n-1) + V_\Theta Y_{ij}(n) \tag{5}$$

Equation (1), (2) and (3) are the mathematical models of the feeding input unit, linking input unit and modulating coupler respectively. Equation (4) and (5) are the expression of the step function and variable threshold function of the pulse generator respectively. Among these equations above, i and j correspond to the location of a neuron in the layer of PCNN, k and l also do.

Parameters denoted in the mathematical models are identical to ones denoted in Fig. 1. M/V_F and W/V_L are the matrix of linking weights/magnitude scaling coefficients of the feedback path and link path respectively. M and W are traditionally local and Gaussian. Θ and V_Θ are the output of variable threshold function and the magnitude scaling coefficient for dynamic threshold respectively. V_Θ is a large constant that is generally more than an order of magnitude greater than the average of U. U is the internal state of the neuron, which will be compared with the dynamic

threshold Θ to produce the output Y, as shown in (4). Each neuron has the same linking strength β in these five equations.

Initially, values of arrays, F, L, U, and Y are all set to zero. The values of the Θ element are initially 0 or some larger value depending on our needs. In the proposed image fusion method using PCNN, the values of the Θ elements are all set to be 0 initially.

2.3 Improved PCNN Model

In Eckhorn's PCNN model, the pulse generator is a step function. In the proposed image fusion method using PCNN, the improved PCNN model is adopted[5, 11]. Equation (4) can be replaced by a novel step function related to the dynamic threshold Θ_{ij}, as shown in (6).

$$Y_{ij}(n) = \begin{cases} \Theta_{ij}(n-1), & \text{if } U_{ij}(n) > \Theta_{ij}(n-1) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The threshold of each PCNN neuron can represent approximately the fire time of the corresponding neuron. A fire mapping image can be obtained by mapping the output threshold of each neuron to the gray level scope of the corresponding image.

The gray level value of the fused image is expressed by the times the neuron fires. Each neuron that has any stimulus will fire in the initial iteration in turn, which will create a large threshold value. It will then take several iterations before the threshold values decay enough to allow the neuron to fire again. It tends to circumvent these initial iterations which contain little information.

3 Experimental Results

In this section, the performance of the proposed image fusion method using PCNN is tested and compared with those of some conventional image fusion schemes using subjective and objective image fusion performance measures. Input image pairs used in the experiments are downloaded from [12]. These image pairs are registered and with 256 gray levels.

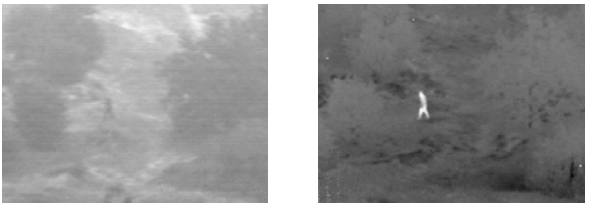
In the experiments, we use three image fusion schemes for comparison, i.e. Principle Component Analysis (abbr. PCA) based image fusion[13], Contrast pyramid based image fusion[14] and Discrete Wavelet Transform (abbr. DWT) based image fusion[15]. In DWT based image fusion scheme, the wavelet decomposition level is 4 and 'symlet 4' are used as wavelets which are compactly supported wavelets and nearly symmetric with highest number of vanishing moments for a given supporting width. The proposed image fusion method using PCNN requires no training. Partial PCNN parameter values used in the experiments are listed as follows[16]:

PCNN iteration time $n = 30$
Delay factor for linking $\alpha_L = 1.0$
Delay factor for dynamic threshold $\alpha_\Theta = 0.012$
Magnitude scaling coefficient of link path $V_L = 0.2$
Magnitude scaling coefficient for dynamic threshold $V_\Theta = 40$

The linking strength β plays an important role in the image fusion method based on PCNN. For a two-channel image fusion method, the value of β for visible image channel is equal to the value of β for infrared image channel. In the succedent experiments, β is set to be 0.5.

Both M and W, the matrix of linking weights of the feedback path and link path, are usually symmetrical, and determined by specific applications.

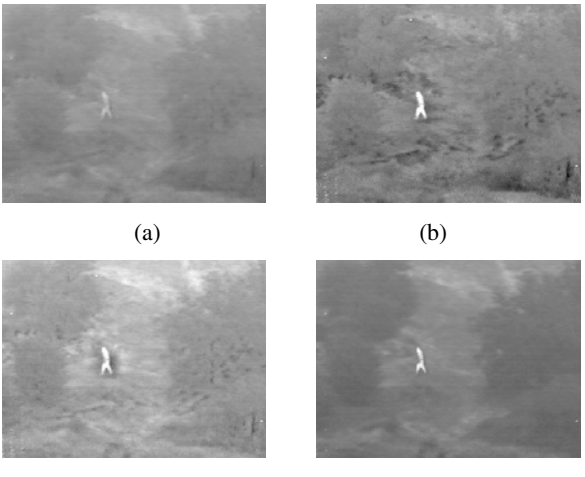
A pair of images is shown in Fig. 3. The image shown in Fig. 3(a) is the visible image, and the image shown in Fig. 3(b) is the infrared image. The important features of the image pair in Fig. 3 are man and trees.



(a) (b)

Fig. 3. Trees.bmp (a) visible image; (b) infrared image

Fused images using different schemes are illustrated respectively in Fig. 4.



(a) (b)

(c) (d)

Fig. 4. Fused images of Trees.bmp (a) PCA; (b) Contrast pyramid; (c) DWT; (d) PCNN

In this experiment, the result of PCA based image fusion algorithm is not very impressive and the outline of the man is blurred to the surroundings. Results of other schemes are as obvious and well-marked as ones in the infrared image. Features both in the visible image and in the infrared one, such as man, trees and middle area, are merged commendably in Fig. 4(d).

Subjective and perceptive comparisons are shown hereinbefore. Then four objective image fusion performance measures are introduced to assess these schemes, i.e. entropy, variance, the fusion performance metric proposed by Xydeas and Petrovic[17, 18] and the fusion performance metric proposed by Piella and Heijmans[19]. These four measures do not require a ground truth or a reference image.

Entropy of an image is a measure of information content[20]. It is the average number of nits needed to quantize the intensities in the image. The more entropy of an image, the more information content is included in the image. It is defined as

$$H = -\sum_{g=0}^{L-1} p(g) \cdot \log_2 p(g) \tag{7}$$

where $p(g)$ is the probability of gray level g, the range of g is [0, …, L-1], and $\sum_{g=0}^{L-1} p(g) = 1$. L is equal to 256 for a gray level image.

Variance, another name for the second moment, measures the gray-level contrast used to establish the relative smoothness[21]. Variance of the image is defined as

$$\sigma^2 = \sum_{g=0}^{L-1} (g - \overline{g})^2 \cdot p(g) \tag{8}$$

where $\overline{g}$ is the statistical mean of gray levels, and it is defined as

$$\overline{g} = \sum_{g=0}^{L-1} g \cdot p(g) \tag{9}$$

The greater variance of an image, the greater gray-level contrast.

The metric proposed by Xydeas and Petrovic can measure the amount of edge information transferred from the source images to the fused image to give an estimation of the performance of fusion algorithms. It uses a Sobel edge operator to calculate the strength and orientation information of each pixel in the source and fused images. Thus overall edge strength and orientation preservation values can be obtained. It is the normalized weighted edge information preservation values $Q^{AB/F}$ that we use to evaluate and compare performance of image fusion algorithms. This performance score falls to the range 0 to 1, with 0 representing total loss of input information and 1 ideal fusion.

The metric proposed by Piella and Heijmans is based on an image quality metric introduced in [22]. In the method, the important edge information of human visual system is taken into account to evaluate the relative amount of salient information contained in each of the input images that has been transferred into the fused image without introducing distortions. The edge-dependent fusion quality index Q_E is calculated with $\alpha = 1$. Similar to $Q^{AB/F}$, this performance score varies from 0 to 1, with 0 representing total loss of input edge information and 1 ideal fusion.

It is noticeable that both Xydeas and Piella metrics are based on edges and consequently fused images containing some significant artifacts can sometimes be inadvertently rated high by the metrics but look inferior perceptually.

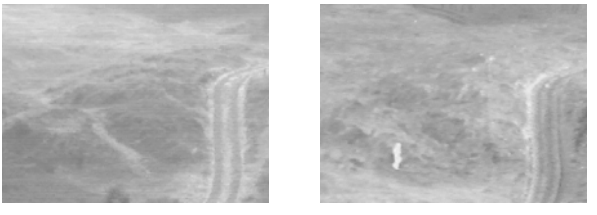
Objective performance assessments of these schemes are shown in Table 1.

Table 1. Performance assessments of image fusion schemes (Trees.bmp)

	PCA	Contrast pyramid	DWT	PCNN
Entropy	5.93	6.18	6.32	5.72
Variance	232.4	350.2	435.1	231.2
$Q^{AB/F}$	0.39	0.52	0.43	0.52
Q_E	0.72	0.73	0.70	0.76

Of all these four image fusion schemes here, $Q^{AB/F}$ and Q_E of the image fused by the PCNN method are the greatest. However, other two objective image fusion performance measures of the PCNN method, such as entropy and variance, are the least than those of other three schemes.

Another pair of images is shown in Fig. 5. The image shown in Fig. 5(a) is the visible image, and the image shown in Fig. 5(b) is the infrared image. The salient features of the image pair in Fig. 5 are man, road and dunes.



(a) (b)

Fig. 5. Dune.bmp (a) visible image; (b) infrared image

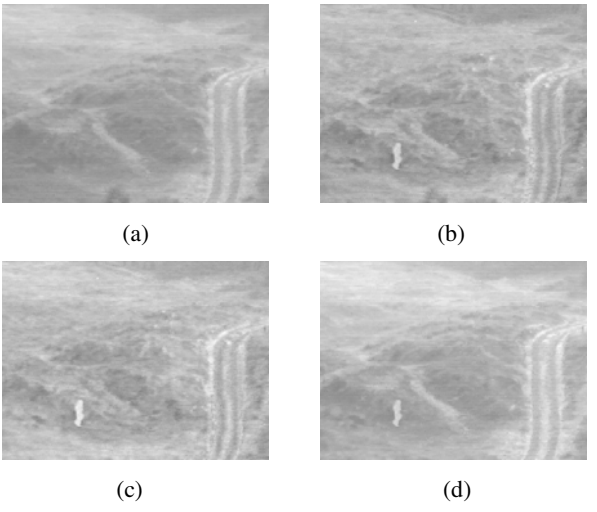
Fused images using different schemes are illustrated respectively in Fig. 6.

In Fig. 6(a), the man is as inconspicuous as in the visible image, despite dunes are clear. The experimental result of PCA based image fusion algorithm shows that it is disabled to fuse the image pair. Except for PCA based image fusion algorithm, other three schemes are effective to merge the visible image and the infrared one. Dunes and the man are clear in the image fused by the PCNN method. It is observable that shapes of dunes are unclear in Fig. 6(c).

Objective performance assessments of these schemes are shown in Table 2.

Of all these four image fusion schemes here, four objective image fusion performance measures of the PCNN method are the greatest in Table 2. These performance assessments of the PCNN method are satisfying.

These four schemes can more or less preserve the salient information and enhance the contrast for visualization. Subjective and objective evaluations show that the PCNN method is effective as a whole.



(a) (b)

(c) (d)

Fig. 6. Fused images of Dune.bmp (a) PCA; (b) Contrast pyramid; (c) DWT; (d) PCNN

Table 2. Performance assessments of image fusion schemes (Dune.bmp)

	PCA	Contrast pyramid	DWT	PCNN
Entropy	6.23	6.06	6.12	6.30
Variance	370.3	266.4	288.3	435.9
$Q^{AB/F}$	0.57	0.50	0.46	0.60
Q_E	0.80	0.75	0.73	0.82

4 Conclusions

In this paper, the image fusion method using PCNN for surveillance is studied. The simplified model and connection model of PCNN neuron are illustrated and the image fusion method using PCNN is presented. The PCNN method is compared with other three image fusion algorithms through some subjective and objective image fusion performance measures. Experimental results show that the PCNN method is effective to merge visible and infrared images from different sensors. Future work will concentrate on a study of automatic determination of PCNN parameters. In addition, because of parallel computation performance of PCNN, a challenging direction for future work could be the extension of the image fusion method using PCNN to the case of real-time image fusion of visible and infrared videos for automatic navigation.

References

1. Eckhorn, R., Reitboeck, H.J., Arndt, M., Dicke, P.: Feature Linking via Synchronization among Distributed Assemblies: Simulation of Results from Cat Cortex. Neural Computation 2, 293–307 (1990)
2. Lindblad, T., Kinser, J.M.: Image Processing Using Pulse-coupled Neural Networks, 2nd edn. Springer, Netherlands (2005)

3. Broussard, R.P., Rogers, S.K., Oxley, M.E., Tarr, G.L.: Physiologically Motivated Image Fusion for Object Detection Using a Pulse Coupled Neural Network. IEEE Transactions on Neural Networks 10, 554–563 (1999)
4. Xu, B., Chen, Z.: A Multisensor Image Fusion Algorithm Based on PCNN. In: Proceedings of the Fifth World Congress on Intelligent Control and Automation, vol. 4, pp. 3679–3682 (2004)
5. Miao, Q., Wang, B.: A Novel Adaptive Multi-focus Image Fusion Algorithm Based on PCNN and Sharpness. In: Proceedings of SPIE, vol. 5778, pp. 704–712 (2005)
6. Wang, Z., Ma, Y.: Dual-channel PCNN and Its Application in the Field of Image Fusion. In: Third International Conference on Natural Computation, vol. 1, pp. 755–759 (2007)
7. Huang, W., Jing, Z.: Multi-focus Image Fusion Using Pulse Coupled Neural Network. Pattern Recognition Letters 28, 1123–1132 (2007)
8. Qu, X.-b., Yan, J.-w., Xiao, H.-z., Zhu, Z.-q.: Image Fusion Algorithm Based on Spatial Frequency-motivated Pulse Coupled Neural Networks in Nonsubsampled Contourlet Transform Domain. Acta Automatica Sinica 34, 1508–1514 (2008)
9. Wang, Z., Ma, Y., Gu, J.: Multi-focus Image Fusion Using PCNN. Pattern Recognition 43, 2003–2016 (2010)
10. Ranganath, H.S., Kuntimad, G.: Iterative Segmentation Using Pulse Coupled Neural Networks. In: Proceedings of SPIE, vol. 2760, pp. 543–554 (1996)
11. Shi, M.-h., Zhang, J.-y., Zhu, X.-j., Zhang, X.-b.: A Method of Image Gauss Noise Filtering Based on PCNN. Computer Applications 22, 1–4 (2002) (in Chinese)
12. Toet, A.: `http://www.imagefusion.org/images/toet2`
13. Chavez Jr., P.S., Sides, S.C., Anderson, J.A.: Comparison of Three Different Methods to Merge Multiresolution and Multispectral Data: Landsat TM and SPOT Panchromatic. Photogrammetric Engineering and Remote Sensing 57, 295–303 (1991)
14. Toet, A., van Ruyven, L.J., Valeton, J.M.: Merging Thermal and Visual Images by a Contrast Pyramid. Optical Engineering 28, 789–792 (1989)
15. Li, H., Manjunath, B.S., Mitra, S.K.: Multisensor Image Fusion Using the Wavelet Transform. Graphical Models and Image Processing 57, 235–245 (1995)
16. Johnson, J.L., Padgett, M.L.: PCNN Models and Applications. IEEE Transactions on Neural Networks 10, 480–498 (1999)
17. Xydeas, C.S., Petrovic, V.: Objective Image Fusion Performance Measure. Electronics Letters 36, 308–309 (2000)
18. Xydeas, C., Petrovic, V.: Objective Pixel-level Image Fusion Performance Measure. In: Proceedings of SPIE, vol. 4051, pp. 89–98 (2000)
19. Piella, G., Heijmans, H.: A New Quality Metric for Image Fusion. In: 2003 International Conference on Image Processing, vol. 3, pp. III-173–III-176 (2003)
20. Hu, L.-m., Gao, J., He, K.-f.: Research on Quality Measures for Image Fusion. Acta Electronica Sinica 32, 218–221 (2004) (in Chinese)
21. Singh, H., Raj, J., Kaur, G., Meitzler, T.: Image Fusion Using Fuzzy Logic and Applications. In: IEEE International Conference on Fuzzy Systems, vol. 1, pp. 337–340 (2004)
22. Wang, Z., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters 9, 81–84 (2002)

Real-Time Performance Reliability Assessment Method Based on Dynamic Probability Model

Cheng Hua[1], Qing Zhang[2], Guanghua Xu[2], and Jun Xie[1]

[1] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, P.R. China
[2] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, P.R. China
huapresent@gmail.com, zhangq@mail.xjtu.edu.cn, xugh@mail.xjtu.edu.cn, jun.xie@stu.xjtu.edu.cn

Abstract. Most previous reliability estimation methods are researched on the assumption of empirical information or prior distribution which is difficult to be acquired in practice. To solve this problem, a real-time reliability assessment method based on Dynamic Probability Model is proposed. The primary step is to establish a Dynamic Probability Model on the basis of nonparametric Parzen window estimating method, and the sliding time-window technique is used to pick statistical samples respectively, then conditional probability density of performance degradation data is estimated. A sequential probability density curve is used to trace the performance degradation process, and probability distribution function on performance degradation data which exceeds the failure threshold is regarded as reliability indicator. Meanwhile, the failure rate is calculated. By analyzing the data from high pressure water descaling pump in the process of failure, it is verified that this method contributes individual equipment to estimate reliability with inadequate empirical information.

Keywords: Dynamic probability model; Reliability estimation; Performance degradation.

1 Introduction

Traditional reliability estimation methods primarily focus on acquiring the failure data of products through a large number of experiments under the same circumstances. Then, the parametric estimation method is utilized based on the selected statistical distribution models according to which the reliability of the products could be predicted. However, these methods lose their scope of application in the event there will be few or no failure for high reliability products in life tests and the acquisition of adequate failure data is impossible in relatively short periods.

In order to eliminate the drawbacks in the traditional reliability estimation methods, a new reliability assessment method which studies the internal relationship between reliability decline and performance degradation is suggested [1-2]. This method enables the track and identification of the degree of reliability decline by establishing a degradation path model of performance variable. Both methods mentioned above focus on the population characteristic of products which is not applied by actual performance

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 88–96, 2010.

degradation process of individual components due to differences in environmental and operational conditions. In 1992, Kin and Kolarik [3] introduced the concept of "real-time reliability" and made attempt to predict the reliability of individual components using real-time performance degradation data. Some real-time reliability assessment methods available today include regression analysis [4-5] and time series analysis [6-7]. Most previous researches are done based on empirical information or prior distribution hypothesis which have imposed restrictions on the promotion and application of the real-time reliability methods, and do not apply to new equipments.

This paper proposes a real-time reliability assessment method based on Dynamic Probability Model. Using this method, a Dynamic Probability Model is constructed to make a real-time estimation of conditional probability distribution of performance degradation data, a sequential probability curve is used to trace the performance degradation process, and probability distribution function which exceeds the failure threshold are regarded as reliability indicator. This method helps to convert performance degradation data into reliability assessment indicator.

The organization of the remainder of this paper is given as follows: real-time reliability estimation method based on time series is introduced in Section 2. In section 3, the dynamic probability model is constructed and illustrated. In section 4, real-time performance reliability assessment method based on Dynamic Probability Model is discussed in detail. In section 5, an actual application case is given to prove this method feasible. Finally, the conclusions are given in Section 6.

2 Overview of Real-Time Reliability Estimation Based on Time Series

The term reliability refers to the ability of a product to perform its required functions under stated conditions for a specified period of time. And whether a product is able to perform specified functions is inferred from whether its performance variable changes within a certain range. In practice, the performance variable which reflects the performance degradation is called degradation variable, and its measured value and true value may be different due to the inevitable measurement error. This measurement can be presented as [8]:

$$z(t) = x(t) + \varepsilon \tag{1}$$

where $x(t)$ is the true value of degradation variable, $z(t)$ is the measured value, ε is the measurement error and $\varepsilon \sim N(0, \sigma^2)$.

In general, the degradation of a product $\{x(t), t \geq 0\}$ can be regarded as a random process, and the distribution of $x(t)$ is presented as:

$$G(x,t) = P(x(t) \leq x) \tag{2}$$

And its probability density function is defined as:

$$g(x,t) = \partial G(x,t) / \partial x \tag{3}$$

The degradation statistical model indicates that a product is considered to be a failure when $x(t)$ reaches the failure threshold L for the first time. Meanwhile, the time when a product fails is presented as:

$$T(l) = \inf\{t : x(t) = l, t \geq 0\} \tag{4}$$

According to the failure time mentioned above, the cumulative failure distribution function of a product can be identified as $F(t \mid L) = P\{T(L) \leq t\}$, which describes the regular pattern of product degradation. Thus it is also known as the model for degradation-failure of a product. The statistical model previous mentioned based on degradation variable and the degradation-failure model indicate an important interrelation between them based on the failure threshold. The statistical inference based on degradation-failure model can be transformed into that based on model of degradation variable [9], which follows the formula:

$$F(t \mid L) = P\{T(L) \leq t\} = P\{x(t) \geq L\} = 1 - G(L,t) \tag{5}$$

Based on this model, reference [6] studies the issue of real-time reliability assessment by analyzing a sequence of performance degradation by time. Now assuming that this sequence of performance degradation is a realization of random process of performance degradation, as shown in Figure 1, at any time t, the performance variable V_t follows normal distribution. Provided that failure threshold V_L is fixed and failure distribution corresponds to the function $F(t,V_L) = P(v(t) > V_L)$, we can get a reliability indicator $R(t) = 1 - F(t,V_L)$, by which the reliability assessment of individual equipment can be accomplished.

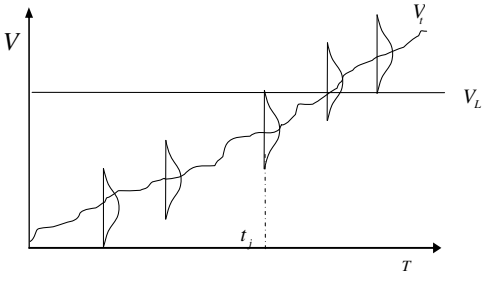


Fig. 1. Reliability assessment curve of performance degradation

The application of reliability assessment based time series is somehow restricted since a large amount of experimental dada or historical information are required to estimate distribution parameter; besides, the application scope of parameter distribution model has some limitations; what is more, the experimental condition is irreproducible. Empirical information of operation equipment is hard to get, and the parameter distribution model does not apply to situations in which operating conditions may be changing. To solve these drawbacks properly, a Dynamic Probability Model (DPM) is introduced into this paper.

3 Dynamic Probability Model

3.1 Principle of Dynamic Probability Model

The model depicted in Figure 2 contains three layers: sampling layer, sample layer and summation layer. Among them the sampling layer consists of equal interval sampling points within the range of observed data, which array orderly. The sample layer contains series of measures data which slide into the network orderly. Summation layer shows the conditional probability density of the set in this situation.

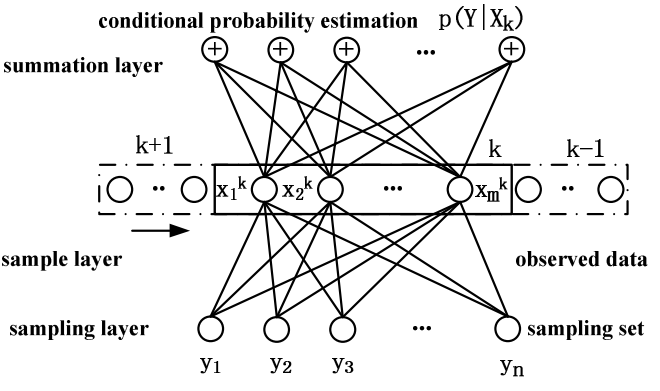


Fig. 2. Dynamic Probability Model

Let X represent a data series. Suppose that the length of the window be m, and the sliding distance be s , whenever the time window slides, $m-s$ observed data are preserved in the window, and the number of newly added data is s. When being computed at the kth time, the samples in the window can be defined as $X_k=\{x_1^k,x_2^k,\cdots,x_m^k\}$. Now let it be mapped to sample layer and after the kernel function operations is done, we can obtain conditional probability function:

$$P(y_i \mid X_k)=\frac{1}{m\sigma}\sum_{j=1}^{m}K(x_j^k,y_i)\quad i=1,2,\cdots,n \tag{6}$$

where σ a smoothing factor ; $K(\cdot)$ is the kernel function and $K(x_j^k,y_i)=\frac{1}{(2m)^{1/2}}e^{-\frac{(x_j^k-y_i)^2}{2\sigma^2}}$ if it is Gaussian function that has been adopted.

When data being observed slide into the time window accordingly, n probability estimation values are got to form a probability distribution curve

$$F(Y)=p(y_i\le Y)=\sum_{\{i:y_i\le Y\}}P(y_i\mid X_k) \tag{7}$$

New probability distribution curve at different time can be get even when new data enters the network by partially updating the estimated probability density value of the previous time.

3.2 The Influences of Parameters in the Mode

There are three parameters in the probability model, namely window length m, sliding distance s and smoothing factor σ, which impact the computation speed and accuracy of the model.

Window length m determines the capacity of the statistical model, sliding distance s determines moving speed of time window. In order to hold consistency of the analysis, m should have a fixed value, and then the same statistical features are granted to every object being analyzed at each moment. S can be adjusted to the condition of objects accordingly to capture the detailed information during the slides. For example, when the condition of the object changes slowly, a relatively longer s can be adopted and vice versa. The selection of the smoothing factor σ is also important: if it is too big, the model will be too smooth to address detailed information in the changing process. If it is too small, the curve will be too sensitive to fake data to tell noise from useful information.

4 Real-Time Reliability Assessment Method

To overcome the limitations which exist in real-time reliability assessment method based on time series, this paper studies two key techniques: the obtaining approach of statistical samples and the dynamic adjustable model of probability distribution.

4.1 Real-Time Acquision of Statistical Samples

Statistical samples used to estimate parameters are hard to obtain without experimental data or historical information. To solve this problem, an approximate method of obtaining statistical samples based on performance degradation data is researched. To reach this goal, the following assumption is proposed:

Assumption: The performance degradation data series is a random time series and the process of performance degradation is nonstationary random process. This process consists of a series of short-time stationary performance degradation data.

Based on the assumption, let a sliding time window ΔT be selected, for the whole degradation period T, $\Delta T / T \to 0$, and the stochastic data series in the window is stationary time series, the statistical feature of which does not change with time, thus the data series in the window can be regarded as statistical sample with identical distribution. At any moment t_k, let data in time window $[t_k - \Delta T, t_k]$ be statistical samples, since $\Delta T / T \to 0$, these samples can also be regarded as the statistical sample at time t_K.

4.2 Dynamic Adjustable Model for Probability Distribution

After the statistical samples are acquired at any moment by using the sliding window technique, the parameter can be estimated according to the assumed parameter

distribution model. Due to the fact that the parameter distribution model depends heavily on samples, the application scope of the model is limited. Even sometimes, the estimated model does not correspond to the actual data. To solve this problem, this paper adopts nonparametric Parzen window estimating method [10], also known as kernel estimation method in which no assumption is necessary. This method gets the asymptotic estimation by weighting local functions in the center of sample points. In theory, this method could approximately approach any density functions. Thus when external circumstances or operation conditions change, dynamic adjustment could be made to adapt changes in actual distribution.

4.3 Real-Time Reliability Assessments

Provided that the real-time acquisition of statistical samples is available, the reliability assessment can be done using failure threshold and the probability distribution of performance degradation data which can be estimated by adopting Dynamic Probability Model in real-time. The process is like the following:

1) Acquire the performance degradation data series ;

2) Select model parameters, namely, window length m , sliding distance s , smooth factor σ and the set of sampling points $Y = \{ y_i \mid i = 1,2,\cdots,n \}$;

3) When performance degradation data enter into the network, a time series is picked using window length m as a statistical sample ;

4) At the kth calculation, data inside the window $V_k = \{ v_1^k, v_2^k, \cdots, v_m^k \}$ are mapped to the sample layer and Gaussian kernel functional calculation is done. Then in the summation layer, conditional probability distribution is got as the probability distribution of performance degradation at time t_K ;

5) Assuming that failure threshold is V_L , by using equation(2) at t_k , the failure probability distribution and performance reliability can be calculated.

The failure probability distribution is

$$F(t_k, V_L) = p(y_i \le V_L) = \sum_{\{i: y_i \le V_L\}} P(y_i \mid V_k) \tag{8}$$

And the performance liability is

$$R(t_k, V_L) = 1 - F(t_k, V_L) \tag{9}$$

And the instantaneous failure rate $\lambda(t)$ is

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T > t)}{\Delta t} = \frac{F'(t)}{R(t)} \tag{10}$$

Since the observed data are discrete, $\lambda(t)$ is calculated through the following difference equation:

$$\lambda(t_k) = \frac{F^{'}(t_k, V_L)}{R(t_k, V_L)} = \frac{R(t_k, V_L) - R(t_{k+1}, V_L)}{R(t_k, V_L)(t_{k+1} - t_k)} \tag{11}$$

6) When $k \leftarrow k+1$, sliding distance of time window is s, the number of observed data being preserved in the window is $m-s$, and the number of newly added data is s, then step 4 and 5 are redone, getting the failure probability distribution of performance degradation and performance reliability at t_{k+1}. As shown in Figure 3, the value of performance reliability at each sampling moment can be achieved by sliding the time-window.

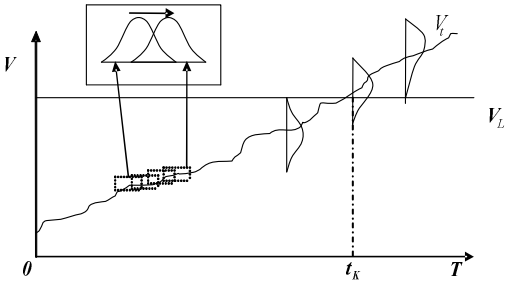
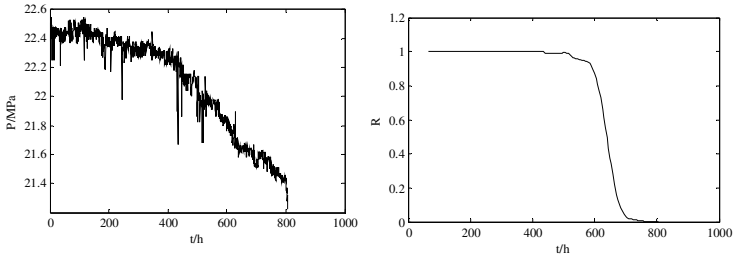


Fig. 3. Tacking curve of performance degradation probability

5 Case Study

The stainless steel plant in Jiuquan Iron and Steel Group Co. (JISCO) uses hydraulic jet descales which are equipped with 11-level centrifugal pump with big flow, which functions to wipe off the oxide scale formed on the surface of stainless steel during the process of steeling rolling. Thus, the pressure at the exist of the pump must meet a certain criteria. That is why pressure sensors are installed at both the entrance and exist of the descaler, and they will record the data in every 30 minutes. For example, the 2# jet descaler in the field was installed in august in 2006, and it has been working since then for 4 months during which 1615 pressure data have been collected. On 13th December, an incident took place on this machine and the cause was later found to be severely worn out of its seal ring. Its worn out process can be regarded as a process of gradual performance degradation, so we can analyze it using the moving neural network. Let the length of sliding window m=120, the distance of sliding s=10 and the smooth factor σ =0.064 ,the number of sampling points n=200, and the failure threshold V_L=21.8MPa (according to the information provided by the manufacturer).

The reliability assessment curve in Figure 4 shows that after being operated for 500 hours, the reliability of the descaler began to decline. And between 600 and 700 hours of operation, its reliability dropped sharply which indicated that its sealing rings were being worn out rapidly. Finally, at 807th hours of operation, an accident happened when the reliability had dropped to 0.



(a) (b)

Fig. 4. Real-time reliability assessment of hydraulic jet descaler: (a) Performance degradation data series; (b) Performance reliability

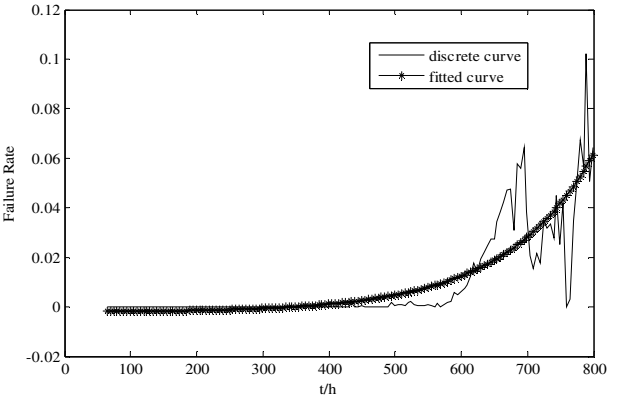


Fig. 5. The failure rates of hydraulic jet descaler

As thown in Figure 5, the dash line is discrete curve of failure rate calculated by function (11). This is an increasing failure rate, known as wear-out failure [11], and the solid line is fitted curve. In this case, the exponential functional form is used to fit discrete failure rate curve by the following formula :

$$\lambda(t) = \alpha + \beta \exp(\phi t) \tag{12}$$

where $\lambda(t)$ denotes failure rate, and α , β , ϕ are constant coefficients of regression model. In this case, the estimation values of regression model parameters are α=-0.0022, β =0.0002, ϕ = 0.0074.

6 Conclusion

The real-time reliability assessment method proposed based on dynamic probability model in this paper manages to calculate the performance degradation data of

equipment without empirical information and distributions being available. The primary step of this work is to establish a dynamic probability model on the basis of nonparametric kernel estimation method, and then sliding time-window technique is used here to pick samples do statistical analysis for the conditional probability distribution. Then the performance degradation data can be converted into reliability assessment indicator by using distribution value of performance degradation data which is more than failure threshold as reliability indicator. Subsequently failure rate is calculated. By analyzing the data from high pressure water descaling pump in the process of failure, this method is proved to be feasible enough to adjust itself to actual running condition and trace the performance degradation process in real-time.

References

1. Carey, M.B., Koenig, R.H.: Reliability assessment based on accelerated degradation: A case study. IEEE Trans. Reliability 40, 499–560 (1991)
2. Lu, C.J., Meeker, W.Q.: Using degradation measures to estimate a time-to-failure distribution. Technometrics 35, 161–174 (1993)
3. Kim, Y.S., Kolarik, W.J.: Real-time conditional reliability prediction from on-line tool performance data. International Journal of Production Research 30, 1831–1844 (1992)
4. Xu, Z., Ji, Y., Zhou, D.: Real-time reliability prediction for a dynamic system based on the hidden degradation process identification. IEEE Transactions on Reliability 57, 230–242 (2008)
5. Gebraeel, N.Z., Lawley, M.A., Li, R., et al.: Residual-life distributions from component degradation signals: a Bayesian approach. IIE Transactions 37, 543–557 (2005)
6. Lu, H., Kolarik, W.J., Lu, S.S.: Real-time performance reliability prediction. IEEE Transactions on Reliability 50, 353–357 (2001)
7. Chan, V., Meeker, W.Q.: Time series modeling of degradation due to outdoor weathering. Communications in Statistics 37, 408–424 (2008)
8. Meeker, W.Q., Escobar, L.A., Lu, C.J.: Accelerated degradation tests modeling and analysis. American Society for Quality 40, 89–99 (1998)
9. Nelson, W.: Analysis of performance-degradation data from accelerated tests. IEEE Trans. Reliability 30, 149–155 (1981)
10. Parzen, E.: On estimation of a probability density function and mode. Annals of Mathematical Statistics, 1065–1076 (1962)
11. Meek, W.Q., Escobar, L.A.: Statitical Methods for Reliability Data. Wiley, New York (1998)

Decomposing Data Mining by a Process-Oriented Execution Plan

Yan Zhang[1], Honghui Li[1], Alexander Wöhrer[2], Peter Brezany[2], and Gang Dai[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University
Shangyuan Residence 3,Haidian Dist, Beijing 100044, China
{07112062,hhli,gdai}@bjtu.edu.cn
[2] Institute of Scientific Computing, University of Vienna
Nordbergstrasse 15/C/3, 1090 Vienna, Austria
{woehrer,brezany}@par.univie.ac.at

Abstract. Data mining deals with the extraction of hidden knowledge from large amounts of data. Nowadays, coarse-grained data mining modules are used. This traditional black box approach focuses on specific algorithm improvements and is not flexible enough to be used for more general optimization and beneficial component reuse. The work presented in this paper elaborates on decomposing data mining tasks as data mining execution process plans which are composed of finer-grained data mining operators. The cost of an operator can be analyzed and provides means for more holistic optimizations. This process-based data mining concept is evaluated via an OGSA-DAI based implementations for association rule mining which show the feasibility of our approach as well as the re-usability of some of the data mining operators.

Keywords: decomposition, data mining operators, data mining execution process plan.

1 Introduction

With the development of information technology, a huge amount of data is accumulated and the scale of data is growing rapidly. For enterprises and governments, data is a kind of wealth which can offer potentials in better policy formulation and improved decision making. Data mining is an effective tool for extracting this useful information from data. Typically, data mining algorithms are black boxes as shown in Fig.1(a). In Weka [8], different data mining algorithms are implemented in components and users can specify one or more options of algorithms through command line or GUI. In Rapidminer [13], the knowledge discovery process is modeled as an operator chain in which the data mining algorithm is just one operator in the chain. In relational database system, the objective of query processing is to transform a high-level query into an efficient execution strategy expressed in a low-level language. A SQL query is rewritten as an execution plan which is composed of basic database operators which can be optimized according to their execution cost. In fact data mining queries are

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 97–106, 2010.

very similar to the SQL query, we can decompose the data mining process in operators of which data mining execution process plan will be composed [16]. This paper continues this research by discussing the cost of operators for association rule- and sequential patterns mining in detail and evaluating an OGSA-DAI [1] based implementation of them. As shown in Fig. 1(b), we will transform data mining algorithm to an execution process plan which is composed of data mining operators.

The rest of paper is structured as follows; Section 2 discusses related work while Section 3 introduces needed background. Section 4 represents the kernel part of this paper by describing the data mining execution process in detail and elaborating on data mining operators for association rules and sequential patterns, including their cost. Section 5 introduces our proof-of-concept implementation together with evaluation results. We finish in Section 6 with conclusions and an outlook on future work.

2 Related Work

Decomposition has been discussed in many data mining literatures, but most of them are based on data set decomposition or in the context of distributed data resources. In the research [10], the feature set and object set decomposition of data mining are introduced, it uses a case to show that the computational complexity and robustness of knowledge extraction from large data sets and decision making can be enhanced by decomposition. [11] summarizes the decomposition methodology in data mining and contains feature set decomposition, sample decomposition, space decomposition and function decomposition. PLANET defines tree learning as a series of distributed computations, and implements each one using the MapReduce model of distributed computation [14]. However, we can see, decomposition in the research mentioned above focus on the different coarse-grained steps of knowledge discovery or is based on the data partitions, what this research focus on is the process of the data mining and the execution process of data mining algorithms.

There are some research about using finer-grained operators in data mining process. [9] proposes Set-oriented mining of association rules in the form of SQL

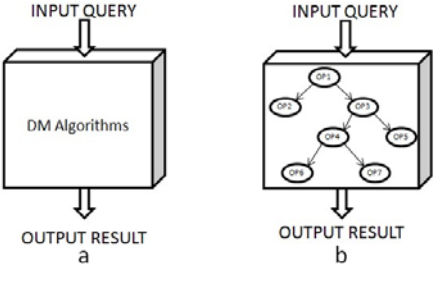


Fig. 1. (a) Traditional black box DM algorithm; (b) DM execution process plan

queries using the SETM algorithm, in which association rules extraction algorithms can be expressed as a set of SQL. [5] presents an algebra for specifying association rule queries where the algebra is based on operators for manipulating nested relations and it allows us to express data mining queries in the same way as ordinary database queries represented by the relational algebra in database systems. [17] uses data mining operators approach to find association rules using nested relations from the database and for this purpose a new query language with the name M2MQL is being defined and semantics of its relevant operator algebra is being discussed. [12] presents the operational semantics of the operator by means of an extended relational algebra. But these research just discuss the operators or steps involved which are not general for other data mining algorithms.

3 Background

The problem of discovering association rules was introduced in [2]. Given a set of transactions D, where each transaction is a set of items, the associations between the body and the head of rule that have support and confidence greater than the user-defined minimum support and minimum confidence respectively are generated, where an association rule is an expression of the form $X \Longrightarrow Y$, where X is the body of the rule and Y is the head of the rule, $X, Y \subset I$, and $X \cap Y = \emptyset, I = \{i_1, i_2, \ldots, i_n\}$ is a set of items and i_j is an item.

[3] first introduced sequential patterns mining. Given a database D of customer transactions, the problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. A sequence is an ordered list of itemsets which can be represented as $S = \langle s_1, s_2, \ldots, s_n \rangle$, where s_i is a set of items, $s_i \subset I$ and the order of itemsets depends upon time or date.

Comparing the association rules mining and sequential patterns mining, we can see both of them need to find frequent patterns from a set of transactions in order to get result rules, the difference lies in the former tries to find frequent itemsets and the latter tries to find frequent sequences, where sequence is an ordered list of itemsets. The phase of finding frequent itemsets should be involved in both data mining processes, so this operator may be reused.

In relational database system, a general outline of the steps required for query processing is shown in Fig.2 [7]. Of course, this sequence is only a general guideline, and different database systems may use different steps or divide one into

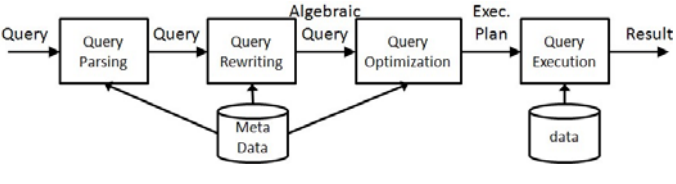


Fig. 2. Query processing steps

multiple steps [6]. In query rewriting phase, the correct calculus query is restructured as an algebraic query, and algebraic queries can be derived from the same calculus query, and that some algebraic queries are "better" than others. The quality of an algebraic query is defined in terms of cost model. The rewritten algebraic query is suitable for subsequent manipulation and optimization.

4 Decomposition of Data Mining Process

The SQL query language is not process oriented and just describes what should be done, while the execution plan is process-oriented and describes how to do things. By reviewing the SQL execution plan, the query author may be able to improve performance of applications and reports with modifications to the SQL statement or the database design. This characteristic is similar with data mining query, so this led us to consider to apply the approach of the execution plan used in relational database system to data mining algorithms. We decompose data mining algorithms into a data mining execution process plan. Such a plan which is process-oriented and composed of data mining operators in order to achieve flexibility, reusability and extendability. The following subsections elaborate on the concerned definitions.

4.1 Data Mining Operators

The granularity of data mining operators presented in this paper is between black box of data mining algorithms and basic database operators used in SQL execution plans. Here are the data mining operators involved in association rules mining and sequential patterns mining. Each operator receives its input, performs specific actions and generates output. All operators take their input in the form of relational tables.We use the following algebra to express these operators, where the subscript (A) means the attributes set, (F) means the formula and (a) means the attribute.

DataScan η Operator: $T_{result} = \eta_{(A),(F)}(T_{source})$. This operator retrieves attributes set (A) and processes attributes according to the formula (F). The input are the source data T_{source}, the specific attributes and formula. The output is the target table T_{result} with attributes (A) and every attribute meets the conditions (F). Here T_{source} and T_{result} can be stored in files or data tables, as well as in the definition of other operators.

FrequentItemsets γ Operator: $T_{result} = \gamma_{(a),(F)}(T_{source})$. This operator generates the set of the frequent itemsets of attribute a and the formula defines the minimum support and minimum confident.

Transformation β Operator: $T_{result} = \beta_{(a)}(T_{source1}, T_{source2})$. This operator replaces every attribute value of $T_{source1}$ with the set of itemsets of $T_{source2}$ which are contained in the attribute value of $T_{source1}$.

ARGeneration α Operator: $T_{result} = \alpha_{(a),(F)}(T_{source1}, T_{source2})$. This operator generates the qualified association rules from the data source $T_{source1}$,

and the data source $T_{source2}$ is used for counting the support and confidence of association rules.

SPGeneration δ Operator: $T_{result} = \delta_{(a),(F)}(T_{source1}, T_{source2})$. This operator generates all qualified sequential patterns from all 1-frequent sequences which is stored in the data source $T_{source1}$ and the data source $T_{source2}$ is used for counting the support and confidence of sequences.

MaximalSeq ζ Operator: $T_{result} = \zeta_{(a)}(T_{source})$. This operator finds the maximal sequences among the set of frequent sequences.

4.2 Data Mining Execution Process Plan

The data mining execution process plan is composed of data mining operators. The operators are combined as directed graph in which each vertex corresponds to a single operator. An arc $a = (X, Y)$ means two operators are connected and the output of the tail operator X should be delivered as input of the head operator Y.

The association rules mining process can be decomposed into the following five data mining operators.

$$T_1 = \eta_{(tid,cid,itemset),(itemsetINNERASC)}(T_{source})$$
$$T_2 = \gamma_{(itemset),(support>msANDconfidence>mc)}(T_1)$$
$$T_3 = \beta_{(itemset)}(T_1, T_2)$$
$$T_4 = \alpha_{(itemset),(support>msANDconfidence>mc)}(T_2, T_3)$$

The relationship between operators is shown in Fig.3. Here, the source transaction table *tblTransactions* consists of three columns: tid (transaction id), cid (customer id), itemset(the items that the customer buys in one shopping). First, DataScan operator projects the columns tid, cid and itemset and sorts the value of column *itemset*. Then FrequentItemsets operator generates frequent itemsets from the column *itemset*. Afterwards, Transformation operator transforms the source table according to the frequent itemsets by replacing the source itemsets with frequent itemsets. Finally, ARGeneration operator generates all qualified association rules from the frequent itemsets.

The sequential pattern mining process can be decomposed into the following six data mining operators.

$$T_1 = \eta_{(tid,cid,itemset),(itemsetINNERASC)}(T_{source})$$
$$T_2 = \gamma_{(itemset),(support>msANDconfidence>mc)}(T_1)$$
$$T_3 = \beta_{(itemset)}(T_1, T_2)$$
$$T_4 = \delta_{(itemset),(support>msANDconfidence>mc)}(T_3, T_2)$$
$$T_5 = \zeta_{(seq)}(T_4)$$

The relationship between six operators is shown in Fig.4. For sequential patterns mining, the source transaction table is same with that in association rules mining. As it is shown in the operators list, the first three operators are same and

it means that these operators are reusable. After frequent itemsets are generated, SPGeneration operator generates all qualified sequences, then MaximalSeq operator filters the maximal sequences.

The above discussion shows the data mining execution plan expressed by the data mining algebra and the directed graph which are easily readable by humans. Internally, XML representation can be used to describe the execution process plan which is readable by machine and is interpreted by the server. The operators are loose coupled in the execution process plan.

4.3 The Complexity of Data Mining Operators

We consider data mining operators as a basis to express the output of data mining query processing. Therefore, the complexity of data mining algebra operations directly affects the execution time of data mining algorithms which can help in choosing the final execution process strategy. Total cost, the sum of all times incurred in processing the operations of the query will be used as the measure of resource consumption [4]. In a centralized database system, the total cost to be minimized includes CPU and I/O costs. The CPU cost is incurred when performing operations on data in main memory. The I/O cost is the time necessary for disk input and output operations. In a distributed database system, the communication cost should be considered which is the time needed for exchanging data between sites participating in the execution of the query [15]. In our research, the communication cost will be considered in optimization phase.

We use the cardinality to define the complexity of operators. Assuming that the transaction data takes the parameters shown in Table 1. We assume that a tuple access is denoted *tupacc* and a tuple computing is denoted *tupcom*. In the association rules mining example, we can estimate which one is the most time consuming operator by analysis the complexity of every operator.

DataScan operator η, the attributes of all tuples are projected and the value of attribute *itemset* are sorted, so the complexity of η is:

$$T \times tupacc + T \times K \log(K) \times tupcom \tag{1}$$

FrequentItemsets operator γ, all non-empty subsets of itemsets of every transaction are generated, and then count the occurrence times of each subsets to get the final frequent itemsets, the complexity of γ is:

$$T \times (2^K - 1) \times tupcom + T \times (2^K - 1) \times \log(T \times (2^K - 1)) \times tupcom \tag{2}$$

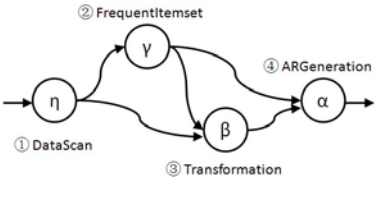


Fig. 3. Association rules mining process

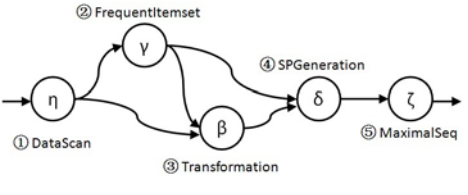


Fig. 4. Sequential pattern mining process

Transformation Operator β, all non-empty subsets of itemsets of every transaction are generated, and then discard the subsets that are not contained in the frequent itemsets, the complexity is:

$$T \times (2^K - 1) \times \log(V) \times tupcom \tag{3}$$

ARGeneration Operator α, all non-empty subsets of every frequent itemset l are generated and for every such subset a, the rules of the form $a \Rightarrow (l - a)$ qualified with the minimal support are final output and all rules are written in a table.

$$V \times (2^{K_V} - 1) \times tupcom + V \times (2^{K_V} - 1) \times tupacc \tag{4}$$

From the above discussion, the complexity of FrequentItemsets operator γ will be the most time consuming, because the number of all non-empty subsets of itemsets of every transaction is exponential order. If a transaction contains 5 items, the number of all non-empty subsets is $2^5 - 1$. The small increase of K can result in the increase of order of magnitude in the number of all non-empty subsets of itemsets. In the next section we will show this from the test.

As we described before, the decomposition can give a clue for the optimization. According to the analysis, we can conclude that the FrequentItemsets operator γ should be the emphasis of the optimization since it is the most time consuming operator. There are many ways about optimization, such as making the execution paralleled based on fragments of data or changing the algorithm about generating frequent itemsets. So after decomposition of data mining as a process-oriented execution plan, we can give a concrete optimization strategy based on single time consuming operator instead of on the whole data mining algorithm.

5 Implementation and Evaluation

In the section we implement the proposed association rules mining execution process plan and evaluate involved data mining operators from different setting of parameters. During the implementation, the proposed data mining execution process plan is very flexible, it can be implemented in central or distributed databases. The clients can build the plan and send the configuration file of the plan to the server through web service, the server is responsible for interpreting and executing the plan by invoking related class instances.

We compare the cost of operators and try to find which operator contributes significantly to the total cost. All transaction dataset used in the experiments are generated by the data generator designed by ourselves. We performed the experiments on a PC with a CPU clock rate of 1.86GHz, 3.9GB of main memory. OS is Fedora9 and the data is stored in MySQL5 database.

We implement the data mining execution plan using OGSA-DAI 3.2 [1], which is a product that allows the sharing of data resources to enable collaboration. It supports data access, transformation, integration and delivery. In our experiments, the data mining operators are implemented as the combination of user defined OGSA-DAI activities. OGSA-DAI and database are installed in the same

Table 1. Parameters of test data

T	Number of transactions
K	Average size of the transactions
N	Number of items
V	Number of frequent itemsets
K_V	Average size of frequent itemsets

Table 2. Parameters settings

Name	T	K	N	minsup
$T1\text{K}{-}K3-N100$	$1K$	3	100	0.02
$T5\text{K}{-}K3-N1000$	$5K$	3	$1K$	0.02
$T10\text{K}{-}K3-N1000$	$10K$	3	$1K$	0.02
$T50\text{K}{-}K3-N10000$	$50K$	3	$10K$	0.02

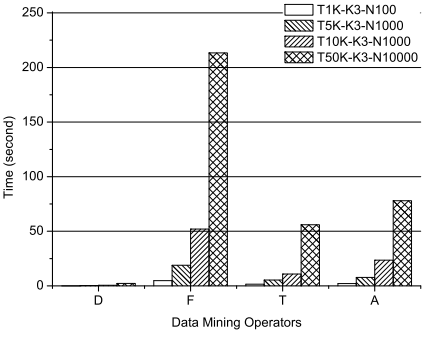


Fig. 5. Ops. perf. for different size of dataset

computer since only centralized data resource is considered in this paper. The MySQL database *DMTransactions* is the data resource registered in OGSA-DAI and the data mining execution process plan is implemented as a workflow and then sent to the engine by web service from the client which is programmed using Java 1.6.

We designed the experiments in order to show the correctness of the analysis in Section 4 and the decomposed data mining process can give users a precise impression of the final system beforehand. The experiments is following:

- Four groups of transaction datasets with different size, as shown in Table2, the K value of each group is 3 and the minimal support is 0.02, different data mining operators performance are shown in Fig.5.
- Four groups of transaction dataset with same size and same minimal support, but the K value is 3, 5 and 7, respectively, the minimal support is 0.02, different data mining operators performance is shown in Fig.6.
- Four groups of transaction dataset with same size and K value, but the minimal support is different, different data mining operators performance is shown in Fig.7.

From the figures we can see that:

- The FrequentItemsets operator contributes significantly to the total cost, especially when the K value gets large.
- The Transformation operator and ARGeneration operator are time consuming when the value of minimal support gets small.

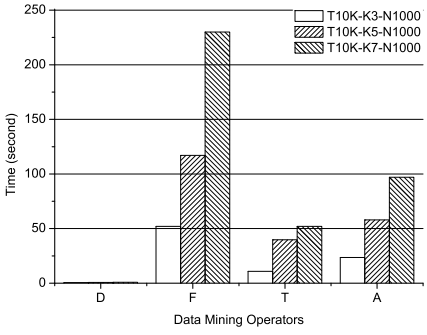


Fig. 6. Ops. Perf. for different K

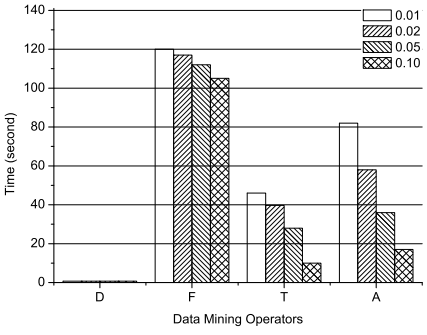


Fig. 7. Ops. Perf. for different min-sup

6 Conclusions and Future Work

Data mining is an effective tool for extracting useful information from large amount of accumulated data which offer potentials in better policy formulation and improved decision making. Traditionally, data mining processes are seen as black boxes implementing some specific algorithms. In order to provide means for reusable, flexible and optimizable data mining processes we elaborated on the decomposition of data mining algorithms into operator based data mining execution process plans.

Our contribution in this work is threefold:

- Operator definitions for two different data mining algorithms, namely association rule mining and sequential pattern mining
- Cost estimations for most relevant operators
- OGSA-DAI based implementation and evaluation showing the feasibility of our approach as well as the re-usability of some of the data mining operators

The decomposition of data mining is the first step of our research and it is the basis of future optimization of data mining process. On the optimization side, we will focus on the optimization of execution process plan in centralized and grid-managed data resources. Additionally, we will implement the proposed data mining execution process plan on China railway freight transport information grid [18] to help extract useful information from distributed waybill data.

Acknowledgments. This work has been supported by Project No. 2006AA01A121 of the National High- Technology Research and Development Program of China. Alexander Wöhrer has been supported by the ADMIRE project which is financed by the European Commission via Framework Program 7 through contract no. FP7-ICT-215024.

References

1. Open grid services architecture - database access and integration (ogsa-dai), http://www.ogsadai.org.uk/
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD 1993, New York, NY, USA, pp. 207–216 (1993)
3. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering, ICDE 1995, pp. 3–14 (1995)
4. Garcia-Molina, H., Widom, J., Ullman, J.D.: Database Systems: The Complete Book, 2nd edn. Prentice Hall, New Jersey (2009)
5. Gopalan, R.P., Nuruddin, T., Sucahyo, Y.G.: Algebraic specification of association rule queries. In: Proceedings of the 4th Data Mining and Knowledge Discovery: Theory, Tools, and Technology (2003)
6. Graefe, G.: Query evaluation techniques for large databases. ACM Comput. Surv. 25(2), 73–169 (1993)
7. Haas, L.M., Freytag, J.C., Lohman, G.M., Pirahesh, H.: Extensible query processing in starburst. SIGMOD Rec. 18(2), 377–388 (1989)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009)
9. Houtsma, M.A.W., Swami, A.N.: Set-oriented mining for association rules in relational databases. In: Proceedings of the Eleventh International Conference on Data Engineering, ICDE 1995, pp. 25–33 (1995)
10. Kusiak, A.: Decomposition in data mining: an industrial case study. IEEE Transactions on Electronics Packaging Manufacturing 23(4), 345–353 (2000)
11. Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook. Springer US, New York (2005)
12. Meo, R., Psaila, G., Ceri, S.: An extension to sql for mining association rules. Data Min. Knowl. Discov. 2(2), 195–224 (1998)
13. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, New York, NY, USA, pp. 935–940 (2006)
14. Panda, B., Herbach, J.S., Basu, S., Bayardo, R.J.: Planet: massively parallel learning of tree ensembles with mapreduce. Proc. VLDB Endow. 2(2), 1426–1437 (2009)
15. Sacco, G.M., Yao, S.B.: Query optimization in distributed data base systems. Advances in Computers 21, 225–273 (1982)
16. Wöhrer, A., Zhang, Y., ul Haq Dar, E., Brezany, P.: Unboxing data mining via decomposition in operators - towards macro optimization and distribution. In: KDIR 2009, pp. 243–248. Funchal-Madeira, Portugal (2009)
17. Yuan, X.: Data mining query language design and implementation. Master's thesis, The Chinese University of Hong Kong, Hong Kong (2003)
18. Zhang, Y., Wöhrer, A., Brezany, P.: Towards China's Railway Freight Transportation Information Grid. In: Proceedings of the 32nd international Convention on Information and Communication Technology, Electronics and Microelectroincs, MIPRO 2009, Opatija, Croatia (2009)

An Efficient Distributed Subgraph Mining Algorithm in Extreme Large Graphs

Bin Wu and YunLong Bai

School of Computer Science Beijing University of Posts and Telecommunications
Beijing 100876, China
wubin@bupt.edu.cn, baiyunl303@163.com

Abstract. Graph mining plays an important part in the researches of data mining, and it is widely used in biology, physics, telecommunications and Internet in recently emerging network science. Subgraph mining is a main task in this area, and it has attracted much interest. However, with the growth of graph datasets, most of these former works which mainly rely on single chip computational capacity, cannot process massive graphs. In this paper, we propose a distributed method in solving subgraph mining problems with the help of MapReduce, which is an efficient method of computing. The candidate subgraphs are reduced efficiently according to the degrees of nodes in graphs. The results of our research show that the algorithm is efficient and scalable, and it is a better solution of subgraph mining in extreme large graphs.

Keywords: subgraph mining; large graph; MapReduce; distributed algorithm.

1 Introduction

Graph-structured data is widely used to represent complicated structures, and it is becoming increasingly abundant in many application domains [4]. The whole society can be represented by a topological graph, in which every person is a node and the relationship of two persons is an edge. Therefore, society can be studied through the research of graph mining. The biologists have found that researches of gene coordination are difficult in biology genetics, which is mainly caused by the changeful reaction between two different genes. A large amount of researches must be done to find out the gene structure's reaction to medicine. However, these kinds of researches are expensive and time consuming. So a befitting method to define the structures of genes is desired. Therefore, the protein structure can be described as a graph structure, in which atoms are the vertices of the figure, while the price of atoms is the corresponding edge. The intrinsic relationship and shared mode between proteins can be dug through graph mining, and these shared patterns can guide the gene experiments in turn. Because of the urgent requirements of these applications, graph mining becomes an important research field, and these researches have brought applications.

In order to discover some useful knowledge from graph-structured databases, graph mining [12], [13], especially subgraph discovery, has been studied intensively. Among most of the applications in graph mining, subgraph is a well-studied problem, since it has a wide and constantly expanding range of applications areas, which include biochemistry, web mining, program flow analysis, etc. Moreover, in the

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 107–115, 2010.

VAST 09 Flitter Mini Challenge the user-defined criminal structure is identified within a social network based on two possible hypotheses. The hypothesis most closely matched in the network represents the criminal social structure [17]. As a consequence, several subgraph mining algorithms have been developed. Some of them rely on principles from inductive logic programming and describe the graph structure by logical expressions [14]. However, the vast majority transfers techniques that were originally developed for frequent item set mining. Examples include MolFea [15], FSG [16], MoSS/MoFa [8], FFSM [9], Gaston [10], gSpan [2] and CloseGraph [5]. Besides those locally optimized algorithms, there are also some parallel or distributed solutions, such as [6]. However, most of these works are based on theoretical distributed methods and do not provide systematic ones. "Cloud" based computing on large scale data has aroused great interests in both industry and research areas [3]. Moreover, recently Google's MapReduce [11] has largely simplified the distributed computing process, without considering any underlying complexity. Meanwhile, MapReduce computational model is also widely applied beyond the cloud as well [1] [4]. Although recent works have focused on graph mining using MapReduce, they are generally quite simple, such as PageRank, or avoiding intractable problems, such as clique enumeration.

In this paper, we provide a parallel subgraph mining algorithm in large graphs. To use the algorithm, the diameter of the motif to be matched must be calculated. Then, we can decide how to represent the large graph. The candidate subgraphs are reduced efficiently according to the degrees of vertexes in graphs. The algorithm is quite efficient and scalable comparing with prior algorithms. This paper is organized as follows: Section 2 presents a short description of preliminaries. Section 3 describes the algorithms which are implemented with MapReduce structure. In section 4, we give an evaluation of our algorithms. Finally, we conclude our work in Section 5.

2 Notation and Definition

Given a graph G, V (G) represents its vertices and E (G) represents its edges. In this paper, we assume without loss of generality that G is simple.

$$\Gamma(v) = \{u \in V(G) \mid (u,v) \in E(G)\}$$

Γ(v) represent the neighbours of v, and P(v, u) represent the shortest path length between v and vertex u.

$$\Gamma_N(v) = \{u \in V(G) \mid P(v,u) \leq N\}$$

$\Gamma_N(v)$ represents the "N-leap" neighbours of v, and $k_v = |\Gamma(v)|$ represents the degree of v.

Definition 1. [Network Diameter] In an undirected graph G, the diameter is the maximum length of all shortest paths. D (G) represents the diameter of G.

Definition 2. [Personal Centre Network] v is a vertex of graph G, the edges between v and its neighbours Γ (v) are the Personal Centre Network of node v. We use C(v) to represent it.

Definition 3. [Local Degree] The degree of a vertex v in a subgraph of G is a local degree of node v. One vertex contains different Local Degree in different subgraph of G. Local Degree is represented by LD(v).

Definition 4. [Subgraph Mining] Suppose that the input database is

$$GD=\{G_i \mid i = 0, 1, 2, \ldots. n\}.$$

If subgraph g is isomorphism with G_i, output G_i.

3 Subgraph Mining Algorithm

First and foremost, we describe the algorithm of subgraph mining when the subgraph's network diameter is 2. Before describing the algorithm, we must represent the large graph as Personal Centre Network. Then we summed up the general parallel algorithm of subgraph mining. All the algorithms are implemented with MapReduce structure.

3.1 Enumerating Personal Centre Networks

We usually represent a graph as the adjacency lists of all nodes, and each adjacency list is the information of a node and its neighbours. Here, in order to dig subgraph, we represent the graph as personal centre network. The following shows a parallel process of transforming representation. To the beginning, we should obtain the "two-leap" information of a specific vertex in the graph, namely v, $\Gamma(v)$ and $\Gamma(\Gamma(v))$ (v's neighbours' neighbours)[1]. Taking the graph in Fig. 1 for example, to find the personal centre network of vertex 1, we can first get < 1, < 2, 3, 4 >>, then < 2, < 3, 4 >> and < 3, 4 > from the adjacency list.

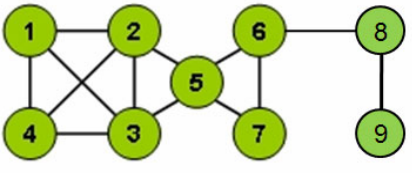


Fig. 1. A simple graph

To reserve the "two-leap" information of a vertex, we present our method, which can also be viewed as subgraph generation based on adjacency list [4]. Let < 1, < 2, 3, 4 >>be a record in the adjacency list, we transform it into the records:

< 2, < 1, <2, 3, 4 >>>
< 3, < 1, < 2, 3, 4 >>>
< 4, < 1, < 2, 3, 4 >>>

For other records in adjacency list, we perform the same operation and finally generate an intermediate result in which each record contains a vertex v, one of its neighbours v' and $\Gamma(v')$. Considering in MapReduce, we present this transformation in a Mapper as in Table 1.

Thus far, given a graph, we transform it into a formation, in which each record conveys a vertex and part of its "two-leap" information. So, we can collect all the "two-leap" information of a vertex and the personal centre network of the vertex can

Table 1. Personal Centre Network

Personal Centre Network	
Map	Input：adjacent list Key：line position Value：<k, $\Gamma(k)$ >
	For each v in $\Gamma(k)$ output <v, <k, $\Gamma(k)$ >> output <k, $\Gamma(k)$ >
	Output：<v, <k, $\Gamma(k)$ >> and <k, $\Gamma(k)$ >
Reduce	Input： Key：v Value：list<v', $\Gamma(v')$ >和 $\Gamma(v)$
	For each vertex vv in $\Gamma(v)$ If vv∈list< v', $\Gamma(v')$ Output <vv , v' >
	Output：Graph record represented as Personal Centre Network

be represented. For example, given a vertex 4, we collect all the records of its "two-leap" records as follows:

< 4, < 1, < 2, 3, 4 >>>
< 4, < 2, < 1, 3, 4, 5 >>>
< 4, < 3, <1, 2, 4, 5 >>>

We can get the Personal Centre Network of vertex 4 is

<4, <<1, 2>, <1, 3>, <2, 3>>>

The whole algorithm is described in Table 1.

3.2 Mining Subgraphs(Whose Diameter Is 2)

Given a certain subgraph named motif as is shown in Figure 2. We can recognize that the network diameter of motif is 2.

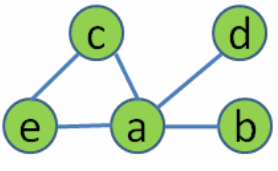


Fig. 2. Motif

At the beginning, we must calculate the subgraph's basic information content, which includes node number and sorted degree list. So, the basic information of motif is as follows: the number of vertex is 5 and the sorted degree list D(m)={4, 2, 2, 1, 1}. We can start the subgraph mining algorithm described below.

First we represent the large graph as Personal Centre Network of every vertex, which is described at section 3.1. Then make the following address of each record (represented as C(v)) of the large graph. We take figure 2 for an example, mining subgraph in figure 1.

Step 1: If the number of vertex in C(v) is less than n, discard C(v). Or, calculate the local degree of all nodes in C(v), and then sort them as a list d(C(v)). For example: the Personal Centre Network of vertex 4 in figure 1 is C(4)=<4, <<1, 2>, <1, 3>, <2, 3>>>, the sorted local degree is LD(C(4))={3, 3, 2, 2}; the Personal Centre Network of vertex 2 is C(2)=<2, <<1, 3>, <1, 4>, <3, 4>, <3, 5>>>, and its sorted local degree is LD(C(2))={4, 4, 3, 3, 2}.

Step 2: According to the sorted local degree list LD(C(v)), obtained by step 1, we select the previous n numbers to compare with D(m). Delete C(v), if LD(C(v))(v) < D(m)(i) ($0 \le i < n$) at the corresponding position. Such as LD(C(4)) and D(m). Otherwise, do the next step.

Step 3: if there has a vertex whose local degree is less than the smallest number in d(m), then delete the vertex and all edges connected to it in C(v).

Step 4: After the first three steps, we calculate all the n combinations of all remaining nodes in C(v). Notice that each of the combination must contain node v. Then, output the combinations and edges between them as candidate subgraphs. For example, the candidate subgraph of C(2) is <2, <<1, 3>,<1, 4>,<3, 4><3, 5>>>.

Step 5: Calculating the candidate subgraph and motif according to graph isomorphism algorithms. We output the candidate subgraphs if they are isomorphic with motif.

We can calculate the sorted local degree list of each candidate subgraph which is the output of Map, and represent the list as LD(c). If LD(c) equals with the motif degree list D(m), continue the next step of the algorithm; otherwise discard the candidate subgraph. Therefore, the efficiency of the algorithm can be improved.

3.3 General Algorithm of Subgraph Mining

In this section, we will give a summary of the general algorithm of subgraph mining. First of all, we should calculate the basic information content of motif, which includes the network diameter D(m), the vertex number N and the sorted degree list L(m). Secondly, transform the representation of large graph to $G_{D(m)}$. Each record $\Gamma_{D(m)}(v)$ of $G_{D(m)}$, which is the "N-leap" neighbours of a vertex v, is processed by the following steps:

Step 1: If the number of vertex in $\Gamma_{D(m)}(v)$ is less than N, discard $\Gamma_{D(m)}(v)$ .Otherwise, calculate the local degree of all nodes in $\Gamma_{D(m)}(v)$. If there has a vertex whose local degree is less than the smallest number in D(m), then delete the vertex and all edges connected with it in $\Gamma_{D(m)}(v)$. The remaining subgraph is represented as $\Gamma'_{D(m)}(v)$.

Table 2. Distributed Subgraph Mining Algorithm

Subgraph Mining	
Map	Input1 : Graph file (graph is represented as "D(m)-leap" of all vertexes) Key : line position Value : $\Gamma_{D(m)}(v)$ =<vertex1, <vertex2, vertex3>…….> Input2 : Basic information of motif. The number of nodes N and the sorted degree list D(m).

	For each vertex V_1 in $\Gamma_{D(m)}(v)$
	If(LD(V_1)<D(m)(N))
	Delete V_1
	For (i=0;i<N;i++)
	If (LD($\Gamma_{D(m)}(v)$)(i) < D(m)(i))
	Delete $\Gamma_{D(m)}(v)$
	If the remaining vertex number of $\Gamma_{D(m)}(v)$ is smaller than N, delete $\Gamma_{D(m)}(v)$. Sorting the local degree of remaining vertexes in $\Gamma_{D(m)}(v)$, representing as LD($\Gamma_{D(m)}(v)$).

	Output: Candidate subgraphs(vertex v, its n-1 neighbors and all edges between them.
Reduce	Input1 : Candidate subgraphs Input2 : motif
	Calculating the candidate subgraph and motif according to graph isomorphism algorithms. output the candidate subgraphs if they are isomorphic with motif
	Subgraphs which are isomorphic with motif

Step 2: Keep on doing step 1 until there is no vertex to be deleted. Then sort the vertexes' local degrees of remaining subgraph as a list LD ($\Gamma'_{D(m)}(v)$).

Step 3: According to the sorted local degree list LD ($\Gamma'_{D(m)}(v)$), obtained by step 2, we select the previous n numbers to compare with D (m). Delete $\Gamma'_{D(m)}(v)$, if

$$\text{LD}\ (\Gamma'_{D(m)}(v))\ (i) < \text{D}\ (\text{m})\ (i)\ (0 \le i < n)$$

at the corresponding position. If the number of remaining vertex is bigger than N, delete $\Gamma'_{D(m)}(v)$. Otherwise, the remaining vertexes and edges between them are represented as $\Gamma^{R}_{D(m)}(v)$, do step 4.

Step 4: After the first three steps, we calculate all the N combinations of the nodes in $\Gamma^{R}_{D(m)}(v)$. Notice that each of the combination must contain node v. Then output the combinations and edges between them as candidate subgraphs.

Step 5: Calculating the candidate subgraph and motif according to graph isomorphism algorithms. We output the candidate subgraphs if they are isomorphic with motif. A complete demonstration of this procedure is described in Table 2.

4 Experiments

All the algorithms are implemented in Java language. The result is obtained on a cluster environment, composed of one master node and 8 slave nodes (Intel(R) Xeon(R) CPU 2.00 GHz, Linux RH4 OS) with 500G*4*8 total storage capacity, and the cluster is deployed with Hadoop-0.20.2 platform. First we test our algorithm by datasets of social network structure in the VAST 09 Flitter Mini Challenge, and the results are the same as what we get in a stand-alone algorithm, though the parallel algorithm is inefficient.

The second experimental dataset is described as follows: it is a service meter datasets of mobile phone comes from one day's communication records in a city in China, and it includes 16,123,172 vertex and 73,636,994 edges. The diameter of the subgraph to be matched is 2 and motif contains 8 vertexes. We use the default Mapper number which is S/64M, S represents input dataset size and 64M is the default store size of each block of the dataset.

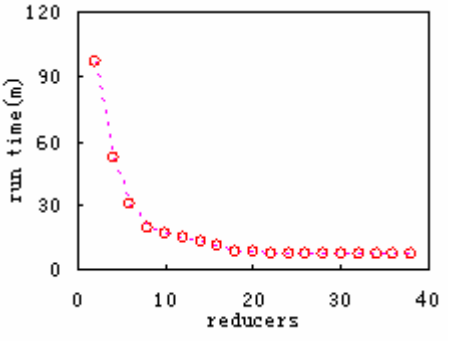


Fig. 3. Runtime with a varying number of reducers

Figure 3 shows the runtime of our algorithms on the dataset (R stands for reducer number). From Figure 3, we observe that the reducing of runtime is not constant as the same times as the increasing of reducer number. Along with increasing reducers' number from 2 to 8, the runtime approximately reduces to 1/2, 1/3 and 1/4. However from the last several number of reducer, we find that along we increase reducers' number from 18 to 64, the runtime does not decrease by times accordingly.

Theoretically the number of reducers can directly indicate the degree of concurrence. Thus the theoretical aggregated throughout versus n reducers is calculated as n×T1, where T1 is the throughout time of the single version [1]. In Figure 4, we plot experimental values in red marks and theoretical values in blue marks. We find the values present a "grow-smooth-decline" procedure. Initially the speedup of our algorithm almost coincides with that in theory. We also find that the optimum speedup is reached at from 8 to 16 reducers, which is just the number of computing nodes (8)

and CPU cores (8 × 2) of our cluster. The speedup deviates from theoretical values after the optimum value significantly. This also shows that after the optimum point, increasing of reducer number cannot get the corresponding efficiencies.

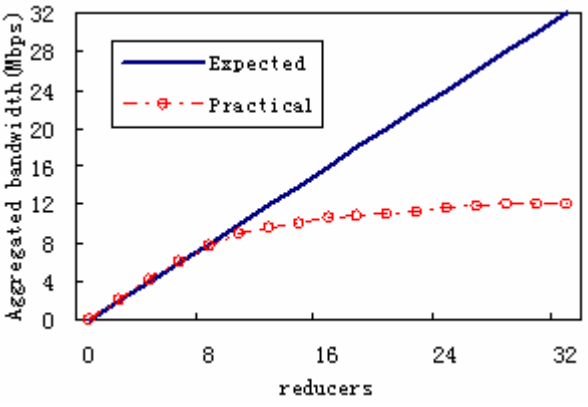


Fig. 4. Performance comparison between expectation and practice

5 Conclusions

In this paper, with the help of cloud-based MapReduce structure, we have developed a parallel algorithm that can perform subgraph mining with a better performance in both theoretical analysis and practical results. In fact, there are a lot of subgraph mining algorithms, but most of these algorithms are inefficient when dealing with extreme large graphs. The algorithm proposed in the paper has the better performance over all other algorithms in large graph mining so far.

The main contribution of the paper is summarized as follows:

1. We presented a massively parallel subgraph mining algorithm, in which the large graph is represented as “n-leap”.
2. Reduce the number of candidate subgraphs by local degrees of the vertexes in each graph record.

Experiments on a massive dataset demonstrate outstanding scalability of the algorithm. Having solved the subgraph mining issue in MapReduce framework, our future work will continue to research on all kinds of graphs in real life, such as directed graph, undirected graph and weighted graph. Besides theoretical methods, we also concern the cloud-based MapReduce applications and make more practical stabs in such powerful tool.

Acknowledgement. This work is supported by the National Science Foundation of China (No.90924029, 6107412860905025), National High Technology Research and Development Program of China (No.2009AA04Z136), National Key Technology R&D Program of China 2006BAH03B05-3, 2006BAJ16B04-02) and the Fundamental Research Funds for the Central Universities.

References

[1] Yang, S., Wang, B., Zhao, H., Wu, B.: Efficient Dense Structure Mining Using MapReduce icdmw. In: IEEE International Conference on Data Mining Workshops, pp. 332–337 (2009)

[2] Yan, X., Han, J.: gSpan: Graph-Based Substructure Pattern Mining. In: Proc. 2nd IEEE Int. Conf. on Data Mining (ICDM 2003), Maebashi, Japan, pp. 721–724. IEEE Press, Piscataway (2002)

[3] Chu, C.T., Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G.R., Ng, A.Y., Olukotun, K.: Map-reduce for machine learning on multicore, pp. 281–288. MIT Press, Cambridge (2006)

[4] Cohen, J.: Graph twiddling in a mapreduce world. Computing in Science and Engineering 11(4), 29–41 (2009)

[5] Yan, X., Han, J.: Closegraph: Mining Closed Frequent Graph Patterns. In: Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD 2003 (2003)

[6] Du, N., Wu, B., Xu, L., Wang, B., Pei, X.: A parallel algorithm for enumerating all maximal cliques in complex network. In: ICDMW 2006, pp. 320–324 (2006)

[7] Kuramochi, M., Karypis, G.: Finding Frequent Patterns in a Large Sparse Graph. In: Proc. 4th SIAM Int. Conf. on Data Mining (SDM 2004), Lake Buena Vista, FL. Society for Industrial and Applied Mathematics, Philadelphia (2004)

[8] Borgelt, C., Berthold, M.R.: Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In: Proc. IEEE Int. Conf. on Data Mining (ICDM 2002), Maebashi, Japan, pp. 51–58. IEEE Press, Piscataway (2002)

[9] Huan, J., Wang, W., Prins, J.: Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism. In: Proc. 3rd IEEE Int. Conf. on Data Mining (ICDM 2003), Melbourne, FL, pp. 549–552. IEEE Press, Piscataway (2003)

[10] Nijssen, S., Kok, J.N.: A Quickstart in Frequent Structure Mining Can Make a Difference. In: Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2004), Seattle, WA, pp. 647–652. ACM Press, New York (2004)

[11] Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. In: OSDI 2004, pp. 137–150 (2004)

[12] Cook, D.J., Holder, L.B. (eds.): Mining Graph Data. Wiley-Interscience, Hoboken (2005)

[13] Washio, T., Motoda, H.: State of the art of graph-based data mining. SIGKDD Explorations 5(1), 59–68 (2003)

[14] Finn, P.W., Muggleton, S., Page, D., Srinivasan, A.: Pharmacore Discovery Using the Inductive Logic Programming System PROGOL. Machine Learning 30(2-3), 241–270 (1998)

[15] Kramer, S., de Raedt, L., Helma, C.: Molecular Feature Mining in HIV Data. In: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2001), San Francisco, CA, pp. 136–143. ACM Press, New York (2001)

[16] Kuramochi, M., Karypis, G.: Frequent Subgraph Discovery. In: Proc. 1st IEEE Int. Conf. on Data Mining (ICDM 2001), San Jose, CA, pp. 313–320. IEEE Press, Piscataway (2001)

[17] Zhou, H., Shaverdian, A.A., Jagadish, H.V., Michailidis, G.: Multiple step social structure analysis with Cytoscape. In: IEEE Symposium on Visual Analytics Science and Technology, VAST 2009, October 12-13, pp. 263–264 (2009)

Spatio-Temporal Clustering of Road Network Data

Tao Cheng[1] and Berk Anbaroglu[1,2]

[1] Dept. of Geomatic Engineering, University College London
Gower Street, London WC1E 6BT, UK
[2] Dept. of Geodesy and Photogrammetry Engineering, Hacettepe University
06800, Beytepe, Ankara, Turkey
{tao.cheng,b.anbaroglu}@ucl.ac.uk,
banbar@hacettepe.edu.tr

Abstract. This paper addresses spatio-temporal clustering of network data where the geometry and structure of the network is assumed to be static but heterogeneous due to the density of links varies cross the network. Road network, telecommunication network and internet are of these type networks. The thematic properties associated with the links of the network are dynamic, such as the flow, speed and journey time are varying in the peak and off-peak hours of a day. Analyzing the patterns of network data in space-time can help the understanding of the complexity of the networks Here a spatio-temporal clustering (STC) algorithm is developed to capture such dynamic patterns by fully exploiting the network characteristics in spatial, temporal and thematic domains. The proposed STC algorithm is tested on a part of London's traffic network to investigate how the clusters overlap on different days.

Keywords: spatio-temporal clustering, road network, spatio-temporal homogeneity and heterogeneity.

1 Introduction

As the amount of data which have spatial and temporal dimensions increases dramatically via the wide usage of sensors and crowd sourcing, spatio-temporal data mining are getting more popular. One of the tasks under spatio-temporal data mining is spatio-temporal clustering, which is the process of grouping data into clusters where the similarity between the observations (an 'observation' in this paper refers to a single measurement which has a value on thematic, spatial and temporal domains) of a cluster is high whereas the dissimilarity between observations at different clusters is high as much as possible. Clustering is used to find the patterns in the data, which will be very useful for pattern analysis.

Among others, finding the non-recurrent traffic congestion is the key activities of Transport for London in order to deliver the journey time reliability for London Olympic Games. However, the dynamics of the traffic reveals the complexity of the network performance which results in traffic congestions change with time, appearing in different size and at different area. The uneven density of road network makes the pattern heterogeneous cross the network. Furthermore, road networks in a week show

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 116–123, 2010.

different patterns due to the travel behaviour difference. Spatio-temporal data include information of three domains which can be used for clustering individually or jointly. Thematic domain defines the characteristics of the data. Spatial domain is used to describe the location of the data. And lastly temporal domain defines the timing of the observation. These domains are used to answer the questions '*what'*, '*where'* and '*when'* respectively.

It is seen that initial research on clustering focused on spatial domain on point data where the density of points are taken into account [8, 10]. Then clustering has been conducted on the combined thematic and spatial domains. [2] combined the spatial and thematic distances into one distance measure by using a pre-defined value (i.e. *w*). This value determines the trade off between the two domains' distances. Choosing *w* is not trivial and it is chosen intuitively. [1] used spatial adjacency relation to cluster on road network data by also considering the traffic flow as the thematic domain. These two domains are combined with each other by considering both the inter-cluster similarity and intra-cluster dissimilarity. [9] mentioned the linkage between detecting spatial clusters and defining the spatial weight matrix. In their research the similarity criteria was based on Getis-Ord local statistic G_i^*. This local statistic is applied for all combinations of the contiguous neighbours and the combination which maximizes this statistic is chosen as the spatial neighbourhood.

Research has also been conducted on the combined temporal and thematic domains. Temporal clustering is the clustering of time series of observed values on the thematic domain. This approach can be used when a retailer wishes to know whether its customers are changing over time or a financial institution wants to determine if the credit card fraud transactions change over time. [3] did temporal clustering by using three indicators; number of clusters that each time stamp will have, minimum number of observations that each cluster should have and the minimum distance between two consecutive clusters; which are defined by the user. Their approach will lead to investigate how clusters at different time stamps are related with each other.

Efforts have been made to detect the clusters considering all three of the domains. For example, [4] divided the time line into fixed size intervals and calculated the similarity based on the thematic domain. A spatial distance threshold is defined to create a graph showing the similarity relations. However, choosing the spatial distance threshold is not a trivial issue. [5] used a probabilistic approach, space-time scan statistics, to detect emerging spatio-temporal clusters. This idea is well suited and applied on disease surveillance. However, the spatio-temporal process is assumed to follow a Poisson distribution which may not be the real case. [7] detected and tracked the clusters of moving point objects. The point observations intersecting in space at some time can be clustered by using the historical clustering information these observations. However, the main limitation is that the number of clusters should be determined a priori.

It is shown that efforts have been made in existing spatio-temporal clustering algorithms to detect clusters changing with time so that clusters capture the dynamicity of spatial phenomenon. Most research, however, either requires threshold values to evaluate the similarities in the spatial or temporal domains, or the number of clusters, which can be tedious for the user. Most importantly, those thresholds or numbers are actually should not be fixed since they are dynamic, relevant to the

thematic attribute of the network. In terms of that, a fully dynamic approach should be adopted for spatio-temporal clustering which will be able to capture the dynamics, homogeneity and heterogeneity in the data.

Next section describes the proposed algorithm to cluster spatio-temporal network data. It is followed by a case study of transport network in London. Conclusion and future work is given in the final section.

2 Proposed Spatio-Temporal Clustering Algorithm

Developed algorithm will exploit the spatial, temporal and thematic domains of the network data. The algorithm is initially proposed at [11] and here it is further developed by incorporating the spatio-temporal search which will be examined under section 2.2. The algorithm is designed for network data and other data types which could be represented as a graph structure (G = (V, E) where V represents the set of spatial objects represented as vertices and E represents the adjacency between the spatial objects. A similar research was done by [6] where the edges are clustered based on their change from presence to absence or vice versa. This paper extends the research at [6] by relaxing the edge's thematic values from binary to real values. The similarity function is based on the values observed at the thematic domain and applied based on the spatial and temporal relations between the objects. These relations define the search direction which the similarity function will be applied.

2.1 Similarity Function

Similarity function is used as the basic component of clustering. It compares two time-series objects (e.g. $p_{1..t}$ and $q_{1..t}$, where p_k and q_k defines the thematic attribute value of objects p and q respectively at k^{th} time and p_k and $q_k \in \mathbb{R}$) which are adjacent to each other (denoted as *adjacent(p,q)*) at their network topology and each having t observations in the temporal domain. Firstly, both of the time-series is divided into equal parts where each part will have only two consecutive observations (p_k, p_{k+1} and q_k, q_{k+1} where k = 1, 2,.., *t-1*). This interval (k to $k+1$) is called as the k^{th} *basic temporal interval (BTI)*. It consists of two consecutive observations in temporal domain so that it is possible to derive several different similarity metrics (slope of change, difference/mean of the two observations,..) to compare between an object and objects which are adjacent to it. Also, all of the possible similarities/dissimilarities between the two compared time series will be captured by this way (since it is not sound to have a basic temporal interval of size one). In other words when comparing the two time series each having t observations, there will be *t-1* similarity comparisons. When two edges are compared by the similarity function the positive similarities will denote the clusters. An intuitive rule for a positive similarity, which shall cover any phenomenon, is that the change (i.e. sign of the slope) between the two consecutive observations should be same at both of the time-series. This condition is formulated as: $\frac{p_k - p_{k+1}}{q_k - q_{k+1}} > 0$.

Apart from the change of direction, it is also believed that the thematic domain values should be close to each other. This closeness is characterized by the δ parameter at the following similarity function. The similarity function forms a linear buffer zone using the sum of the two consecutive observations and if the other time series' sum of the two consecutive observations falls in that buffer zone, then the time-series' exhibits a positive similarity. For simplicity, at further references to the similarity function, δ and *adjacent(p,q)* parts will be omitted.

$$simF(p_k, p_{k+1}, q_k, q_{k+1}, \delta \mid 0 < \delta < 1; p, q \in V \wedge adjacent(p,q)) = \begin{Bmatrix} 1, if \left| \frac{p_k + p_{k+1}}{q_k + q_{k+1}} - 1 \right| < \delta \\ 0, otherwise \end{Bmatrix}. \quad (1)$$

Yet, this is not the only similarity function that can be defined. Other binary similarity functions, as long as they take two consecutive observations of the two time series, can be defined by using background knowledge.

2.2 Search Directions

Once the similarity function is determined, the algorithm searches for positive similarities to form spatio-temporal clusters. Searching in spatio-temporal domain can be investigated under three categories: spatial-same time, temporal-same space and spatio-temporal.

Spatial-same time search compares two adjacent objects at same times. In other words, similarity function is evaluated for all of the adjacent objects (i.e. for all p and q pairs that are *adjacent(p,q)*): $simF(p_k, p_{k+1}, q_k, q_{k+1} \mid k = 1,2,..,t-1)$. Similarity search continues with the objects that are adjacent to q at the positive similarities found at the comparison of p and q. This recursive search continues until there are no more positive similarities found.

Second search direction is temporal-same space which analyses the temporal similarity in between the objects themselves at consecutive times: $simF(p_k, p_{k+1}, p_{k+1}, p_{k+2} \mid k = 1,2,..,t-2)$. This search direction is expected to give the highest number of clusters and there is no need for recursive search.

Third search direction, spatio-temporal search, is the search conducted in both space and time. The main idea in this search lies to model the flow in the network. Since in the road network vehicles flow, an event happening at one edge might show its effect at its adjacency after a time elapse. The main question under this search direction is to find the number of time steps which quantifies this time elapse. This number is referred as *stepSize*. The similarity function that will be used will be like: $simF(p_k, p_{k+1}, q_{k+stepSize}, q_{k+stepSize+1} \mid k = 1,2,..,t-stepSize-1)$.

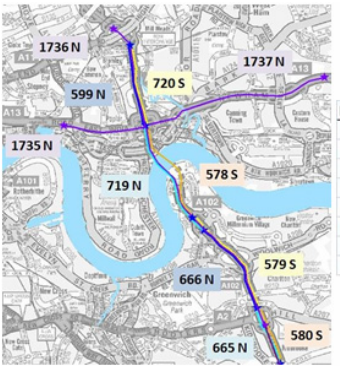
3 Case Study –Blackwall Tunnel at London

3.1 Data Description

Data is collected via the automatic number plate recognition (ANPR) cameras at 5 minute intervals between 28 December 2009 – 3 January 2010 for 11 links (a link is

defined as the road segment(s) where two ANPR cameras –one at the beginning and one at the end of the link- operate and collect data according the number of cars passed) at a region which is renowned with its congestion; Blackwall Tunnel. This region has a road that passes underneath the river Thames and the region is very close to the banking centre of London; Canary Wharf. Spatial dimension is used via the adjacency matrix, temporal dimension can be thought in different scales, as minute, day or week based and the thematic domain is the average speed in km/h. Aim of the case study is to observe how spatio-temporal clusters occurring on different days of a week differ from each other and analyze how adjacent links behave.

There are 11 links where 3 of them (i.e. 1735N, 1736N, 1737N) overlap with the rest of the links. Because of this overlap, they are not adjacent to any of the links and discarded from the analysis. Letters shown near the link numbers denote the direction of flow, where N and S mean that the traffic on that link flows towards north and south respectively. The study area and the adjacency matrix is illustrated at figure 1.



links	578	579	580	599	665	666	719	720
578	0	1	0	0	0	0	0	1
579	1	0	1	0	0	0	0	0
580	0	1	0	0	0	0	0	0
599	0	0	0	0	0	0	1	0
665	0	0	0	0	0	1	0	0
666	0	0	0	0	1	0	1	0
719	0	0	0	1	0	1	0	0
720	1	0	0	0	0	0	0	0

Fig. 1. Blackwall Tunnel Region and the Adjacency Matrix

It can be seen that the traffic flow direction is incorporated in defining the adjacency matrix. If the directionality of the links has not been incorporated then for example 579 and 666 will be adjacent, however they are not adjacent since 579th link is flowing towards south whereas 666th links flows towards north.

Two parameters should be determined before searching for detecting the clusters. First is the δ parameter which will be used at the similarity function stated at equation 1. It is chosen as 0.1 after consulting to the colleagues at Transport for London. Second parameter is the stepSize, which is chosen after investigating the average journey time for the links on the first day of the analysis. The average journey time for the links is shown at figure 2. Solid black line shows the 5 minute level, and it is seen that most of the journeys are completed less than 5 minutes. This suggests that stepSize can not be not greater than 5 minutes, thus stepSize is chosen as one BTI. In other words, when the spatio-temporal search is conducted the similarity function will be in the form of $simF(p_k, p_{k+1}, q_{k+1}, q_{k+2} | k = 1,2,..,t-2)$.

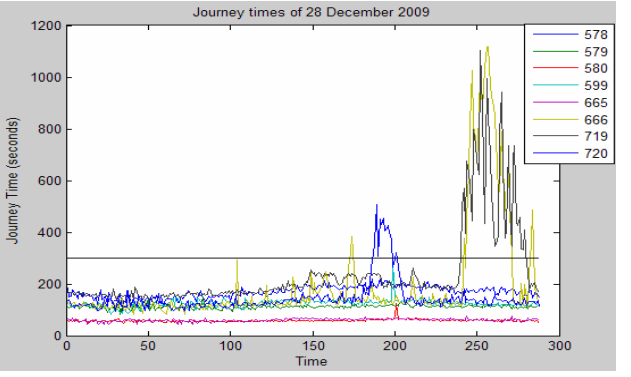


Fig. 2. Journey times of 28 December 2009 at Blackwall Tunnel Links

3.2 Results

Two of the search directions, spatial-same time and spatio-temporal, are tested since temporal-same space search is pure temporal analysis and does not reveal information about the network as a whole.

Each detected cluster has two components. First component states the links involved in that cluster and second denotes the BTIs which those links are clustered. This representation does not reveal much information, thus the representation is transformed into a binary time series where; if there is a cluster at the BTI, the function outputs one otherwise zero. These binary sequences of clusters were suitable for comparing the clusters that are detected on different days. To compare these binary sequences, Jaccard Coefficient (JC) is used, which is defined as:

$$JC = \frac{n_{11}}{n_{10}+n_{01}+n_{11}}. \quad (2)$$

where n_{11} denotes the number of times when both of the days has a cluster at that BTI (i.e. positive match), n_{00} denotes the number of times when both of the days does not have a cluster at a BTI (i.e. negative match) n_{10} and n_{01} denotes the number of mismatches where one day has a cluster at a BTI whereas there is no cluster at that BTI in the other day.

For experimental purposes, clusters that occur on the first day of the analysis (28 December 2009) is compared with the clusters that occurred at the rest of the days (29 December – 3 January). It can be verified from the JCs that the days show distinct characteristics.

Spatial-same time and spatio-temporal search directions will be done in order to capture the spatio-temporal clusters.

Table 1 shows the JCs between the clusters shown at the left column and at comparison between the first day and the other days (day 2 –day 7) on the spatial-same time based search. '[ ]' means that either of the days does not have that cluster. It is seen that links 578 and 579 has a cluster only at the first and second days and at the other days; there is no 578-579 cluster which strongly suggests that these two links have different characteristics.

Table 1. Jaccard Coefficients for the spatial-same time search

Clustered Links	Day2	Day3	Day4	Day5	Day6	Day7
[578,579]	0	[]	[]	[]	[]	[]
[578,720]	0.2	0.18	0.24	0.2	0.19	0.25
[579,580]	0.2	0.19	0.28	0.23	0.25	0.24
[599,719]	0.17	0.12	0.12	0.14	0.13	0.13
[599,719,666]	0.07	0	0	0	0.04	0
[665,666]	0	0.13	0	0	0.17	0.1
[666,719]	0.14	0.07	0.1	0.13	0.15	0.06

Similar idea is applied when the spatio-temporal search is conducted and the Jaccard Coefficients are shown at table 2. The previous results are validated as there is no relation between the links 578 and 579 in spatio-temporal search as well. In addition, the links 579 and 580 showed the highest overlap between the days.

Table 2. Jaccard Coefficients for the spatio-temporal time search

Clustered Links	Day2	Day3	Day4	Day5	Day6	Day7
[578,720]	0.20	0.21	0.17	0.20	0.17	0.16
[579, 578]	0	[]	[]	[]	[]	0
[579, 580]	0.24	0.27	0.24	0.24	0.23	0.25
[580, 579]	0.29	0.19	0.2	0.21	0.24	0.2
[599, 719]	0.19	0.1	0.11	0.1	0.08	0.1
[599, 719, 666]	0.05	0	0	0	0	0.03
[665, 666]	0	0	0	0.1	0	0.14
[665, 666, 719]	0	0	0	[]	[]	[]
[666, 665]	0	0	0	0	0.09	0.06
[666, 719]	0.07	0.06	0.11	0.05	0.13	0.11
[666, 719, 599]	0.05	0.03	0	0	0	0
[719, 599]	0.15	0.15	0.09	0.14	0.12	0.11
[719, 666]	0.08	0.12	0.16	0.16	0.08	0.06
[720, 578]	0.24	0.16	0.24	0.2	0.28	0.2

From these results it is seen that daily patterns vary, and should be treated accordingly. In addition, number of positive similarities decrease as the spatial search is extended; which is indeed an expected outcome.

4 Conclusion

This research proposed an algorithm of spatio-temporal network clustering where spatial and temporal domains are exploited using the network topology (via adjacency matrix) and characteristics (via spatial-same time and spatio-temporal search directions). It is shown that, the proposed algorithm captures the dynamics of the network so that the spatio-temporal clusters appear and disappear with time. In addition, by this way both homogeneity and heterogeneity in the data is captured. Finally, user is involved only in determining the similarity on thematic domain which can be decided by domain expert.

Future work will be focused on different scales in temporal domain. In most the researches, as well as the case study mentioned in this paper, temporal dimension exhibits in different scales (time-of-day, day-of-week, etc.) and scaling the results at these different scales is a challenging problem. Secondly, detecting the appropriate

stepSize is also a challenging problem, since this research assumed that stepSize does not change spatio-temporally. Finally, the linkage between the spatio-temporal clusters and the traffic congestion will be sought.

Acknowledgments. This research is jointly supported by UK EPSRC (EP/G023212/1), Chinese NSF (40830530) and 863 Programme (2009AA12Z206). Thanks to Transport for London for providing the data and domain knowledge. Second author thanks to Hacettepe University and Higher Education Council of Turkey for the PhD scholarship they provide.

References

1. Wang, Y., Chen, Y., Qin, M., Zhu, Y.: SPANBRE: An Efficient Hierarchical Clustering Algorithm for Spatial Data with Neighborhood Relations. In: Fourth International Conference on Fuzzy Systems and Knowledge Discovery. FSKD 2007, pp. 665–669 (2007)
2. Lin, C., Liu, K., Chen, M.: Dual clustering: integrating data clustering over optimization and constraint domains. IEEE Transactions on Knowledge and Data Engineering 17, 628–637 (2005)
3. Adomavicius, G., Bockstedt, J.: C-TREND: Temporal Cluster Graphs for Identifying and Visualizing Trends in Multiattribute Transactional Data. IEEE Transactions on Knowledge and Data Engineering 20, 721–735 (2008)
4. Wei, L., Peng, W.: Clustering Data Streams in Optimization and Geography Domains. Advances in Knowledge Discovery and Data Mining, 997–1005 (2009)
5. Neill, D.B., Moore, A.W., Sabhnani, M., Daniel, K.: Detection of emerging space-time clusters. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 218–227. ACM, Chicago (2005)
6. Chan, J., Bailey, J., Leckie, C.: Discovering correlated spatio-temporal changes in evolving graphs. Knowledge and Information Systems 16, 53–96 (2008)
7. Rosswog, J., Ghose, K.: Detecting and Tracking Spatio-temporal Clusters with Adaptive History Filtering. In: IEEE International Conference on Data Mining Workshops. ICDMW 2008, pp. 448–457 (2008)
8. Yiu, M.L., Mamoulis, N.: Clustering objects on a spatial network. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, pp. 443–454. ACM, Paris (2004)
9. Aldstadt, J., Getis, A.: Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. Geographical Analysis 38, 327–343 (2006)
10. Harel, D., Koren, Y.: Clustering spatial data using random walks. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 281–286. ACM, San Francisco (2001)
11. Cheng, T., Anbaroglu, B.: Defining Spatio-Temporal Neighbourhood of Network Data. In: ISGIS, pp. 75–80 (2010)

A Sampling Based Algorithm for Finding Association Rules from Uncertain Data

Zhu Qian, Pan Donghua, and Yang Guangfei

Institute Systems Engineering, Dalian University of Technology,
Dalian, China
zhuqianyx@gmail.com, gyise@dlut.edu.cn, yangguangfei@gmail.com

Abstract. Since there are many real-life situations in which people are uncertain about the content of transactions, association rule mining with uncertain data is in demand. Most of these studies focus on the improvement of classical algorithms for frequent itemsets mining. To obtain a tradeoff between the accuracy and computation time, in this paper we introduces an efficient algorithm for finding association rules from uncertain data with sampling-SARMUT, which is based on the FAST algorithm introduced by Chen et al. Unlike FAST, SARMUT is designed for uncertain data mining. In response to the special characteristics of uncertainty, we propose a new definition of "distance" as a measure to pick representative transactions. To evaluate its performance and accuracy, a comparison against the natural extension of FAST is performed using synthetic datasets. The experimental results show that the proposed sampling algorithm SARMUT outperforms FAST algorithm, and achieves up to 97% accuracy in some cases.

Keywords: uncertain data, association rule mining, sampling; data mining.

1 Introduction

Data mining techniques are heavily used to search for relationships and information that would be "hidden" in transaction data. Association rule mining is one of the most important data mining techniques and a means of identifying relationships among sets of items. An important step in the mining process is the extraction of frequent itemsets, or sets of items that co-occur in a major fraction of the transactions [1,2].

In the databases of precise data, users definitely know whether an item is present in a transaction in the databases. However, there are many situations in which users are uncertain about the presence of some items [3]. For instance, a medical dataset may contain a table of patient records, each of which contains a set of symptoms that a patient suffers. Applying association analysis on such a dataset allows us to discover any potential correlations among the symptoms and illnesses. In many cases, symptoms of subjective observations would best be represented by probabilities that indicate their presence in the patients' tuples [4].

These real-life situations in which data are uncertain call for efficient mining algorithms from uncertain data. This problem has been studied in a limited way in [1,3,5], and a variety of pruning strategies are proposed in order to speed up the

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 124–131, 2010.

algorithm. Agarwal, C.C., et al. [6] examined the behavior of the extensions of well known classes of deterministic algorithms and found that the extensions of the candidate generate-and-test as well as the hyper-structure based algorithms are more effective than the pattern-growth algorithms.

However, there are still no new efficient methods proposed for frequent pattern mining from uncertain data so far. Since the inclusion of probability information is that the size of the dataset would be much larger than that under the quantized binary model, which makes the mining consume more time and memory usage. One popular approach is to run mining algorithms on a small subset of the data instead of the entire database, sometimes called sampling, which is often possible to explicitly tradeoff processing speed and accuracy of results [7]. Some sampling methods work very well and significantly reduce the time for association rule mining in precise data, but not uncertain data. Based on this, we investigate sampling methods that are designed to deal with mining algorithms for uncertain datasets.

In this paper, we propose an efficient algorithm based on sampling for association rules mining in existential uncertain transactional datasets called SARMUT, which is a modification and extending of FAST (Finding Association rules from Sampled Transactions), a classic algorithm that was presented in [8]. Through experiments, we will show that SARMUT is very efficient in terms of both CPU cost and accuracy, even performs better than the natural extension of FAST.

This paper is organized as follows. The next section gives related work and background. In Section 3, we introduce our SARMUT sample algorithm for mining frequent patterns from uncertain data. Section 4 shows experimental results that compare with FAST. Finally, conclusions are presented in Section 5.

2 Related Work and Background

2.1 Frequent Itemsets Mining Algorithm

Agrawal [9] proposed the Apriori algorithm using generate-and-test strategy to find all frequent patterns from a transaction database containing precise data. Each transaction in the database consists of items which all have a 100% probability of being present in that transaction.To improve efficiency of the mining process, Han et al.[10] proposed an alternative framework, namely a tree-based framework, to find frequent patterns.

2.2 Uncertain Data

A key difference between precise and uncertain data is that each transaction of the latter contains items and their existential probabilities. The existential probability $P(x,t_i)$ of an item x in a transaction t_i indicates the probability of x being present in t_i. There are many items in each of the $|D|$ transactions in a transactional database D and x is items in an itemset X. Hence, the expected support of X, denoted by exp$Sup(X)$ can be computed as follows:

$$\exp Sup(X) = \sum_{i=1}^{|D|} \left(\prod_{x \in t_i} P(x, t_i) \right). \quad (1)$$

2.3 Sampling-Based Association Rule Mining

Sampling is a powerful data reduction technique that has been applied to a variety of problems in database systems. Zaki, et al. [12] employed a simple random sample to identify association rules and Toivonen [13] uses sampling to generate candidate itemsets but still requires a full database scan.

A classical algorithm-FAST (Finding Associations from Sampled Transactions), is introduced by Chen, et al. [8]. The novel two-phase strategy to collect samples of FAST is different from previous sampling-based algorithms. In the first Phase, a large initial sample of transactions is collected and used to estimate the support of each item in the database. In the second Phase, a small final sample is obtained from the initial sample so that the final sample is as close as possible to the original dataset.

3 Our Proposed SARMUT Algorithm

In this section, we will describe the SARMUT algorithm in details. First, we introduce the basic algorithm and some definitions we will use.

3.1 Overview of the Algorithm

Notation 1. Denote by D the original database and by S a simple random sample drawn without replacement from D. Let $|D|$ represent the number of transactions in D and $|S|$ represent the number of transactions in S. Let $I(D)$ be the collection of itemsets that appear in D; a set of items X is an element of $I(D)$ if and only if the items in X appear jointly in at least one transaction $t \in D$. For a set of transactions T, let $\exp Sup(X,T)$ be the expected support (computed according to Equation (1)) of X in T. Then the support of X in S is given by $\sup(X,S) = \exp Sup(X)/|S|$.

Given a specified minimum support $\min Sup$, the SARMUT algorithm proceeds as follows:

1. Extract a relatively larger simple random sample S from D;
2. Compute $\sup(X,S)$ for each 1-itemset and the averages of all transactions;
3. Using the supports and sum computed in Step 2, trim outliers from S to obtain the final small sample $S0$;
4. Implement a standard association rules mining algorithm against $S0$—with minimum support $\min Sup$—to obtain the final set of association rules.

In Step 1, the random sampling approach we use performs without replacement. Steps 2 is straightforward and step 4 uses a standard association rule mining algorithm such as UApriori which is an extension of set-enumeration algorithm proposed in [6].

The crux of the algorithm is in step 3. Distance is denoted by the measure of difference between two transactional sets to choose which transactions should be trimmed away, which will be described in details in the following subsection.

3.2 The Distance Function

Unlike FAST which is designed just for precise count data applications, our proposed SARMUT is modified to fit for the association rules mining in uncertain data. The key modifications to FAST are as follows:

1. The method of expected support computation is changed to fit for uncertain data and it uses Equation (1) to calculate the value of each frequent itemset.

2. The measure used to sort out transactions contained in the final sample is modified in SARMUT. We add another characteristic, the average of all existential probabilities in a transaction to the definition of "distance".

Since the final sample has a chance of losing some information of the original database, a good sampling algorithm is able to select a collection of transactions that is very similar to the whole database in order to produce the same frequent patterns mining result. We use "distance" as a measure that compares the difference between the sample set and the whole set. In this context, the definition of the measure "distance" is very important to the effectiveness of the algorithm. The measure "distance" based on the characteristics of transaction database should be able to reflect useful information in the whole set. Chen, et al [8] used the difference with respect to the expected supports of all 1-itemsets to define "distance". In view of the property of transactional database with uncertain data, we add another characteristic to measure the distance between the sample set and the whole set.

According to Equation (1), we can deduce that the larger $P(x,t_i)$ is, the larger exp$Sup(X)$ ($x \in X$) is. Larger existential probabilities exert more remarkable influence on the value of the expected supports. It is necessary to select typical transactions with larger existential probabilities to the sample set. Based on this conclusion, we add the average of all existential probabilities in a transaction to the "distance", denoted by $T\_dist(t_i)$, which can be computed as follows:

$$ave(t_i) = \sum_{x \in t_i} P(x, t_i) / t_i\_size\ , \tag{2}$$

$$T\_dist(t_i) = \left(ave(t_i) - \sum_{j=1 \& j \neq i}^{|Sm|} ave(t_j) \right) / (|Sm| - 1)\ . \tag{3}$$

Let t_i_size be the number of items in transaction i, and t_i represent transaction i. Let S be the initial sample set, Sf be the final sample set and Sm be the current sample which is the subset of S. The discrepancy of Sm and S is computed by using the I_dist metric:

$$I\_dist(Sm, S) = \sum_{X \in I1(S)} \left| \sup(X, Sm) - \sup(X, S) \right|\ . \tag{4}$$

Equation (3) and (4) represent the vertical and horizontal characteristics of the database respectively. Then we merge these two characteristics into one equation:

$$Dist(Sm, S) = I\_dist(Sm, S) + 0.01 \times i\_size \times T\_dist(t_i)\ . \tag{5}$$

where $t_i = \{S\text{-}Sm\}$.

3.3 The Trimming Method

Given a desired number of transactions in the final sample *final_size*, our goal is to choose *fianl_size* transactions to constitute the final sample in order to make the *Dist*(*Sm,S*) (|*Sm*|=|*Sf*|) minimal. The problem can also be represented as follow:

$$\text{minimize } Dist(Sm, S) . \tag{6}$$

We can see that this combinatorial optimization problem is NP-complete. Since the database constituted by transactions can be broken into some small parts, we can use an alternative greedy algorithm to find an approximate solution. A greedy algorithm follows the problem solving metaheuristic of making the locally optimal choice at each stage with the hope of finding the global optimum. The defect of such methods is their high calculation burden in which case the quantity of data is very large. Therefore in SARMUT, the goal is to explicitly trade off speed and accuracy.

```
Extract a larger simple random sample S from D;
Set Sm = S;
Compute sup (X,S) for each 1-itemset X∈I1 (S);
Compute the sum of all ave(ti) for each transaction ti in Sm;
while(|Sm| >fianl_size)
{  Divide Sm into disjoint groups of k transactions each;
    for each group G
    {
     Compute sup(X,Sm) for each item X in Sm;
     Set Sm = Sm −{t*},where
          Dist(Sm −{t*},S)=min ti∈G Dist(Sm −{ ti },S);
    }
}
Implement a standard association rules mining algorithm against Sf —with
minimum support minSup;
Obtain the final set of association rules.
```

Fig. 1. The basic SARMUT Algorithm

By choosing a value of k between 1 and |S|, the algorithm starts with setting Sm=S. Then divide Sm into groups and each group contains k transactions. For each group, we compute every *Dist*(*Sm,S*) where Sm=$\{S$-$t_i\}$ and $t_i \in \{t_1, t_2,..., t_k\}$ in the group denoted by G. The minimal *Dist*(*Sm,S*) will be found and the corresponding transaction t_i will be trimmed from the sample Sm, because its removal will result in the slightest variation in Equation (5). Repeat this procedure until the size of Sm reaches to *final_size* given by users. The basic SARMUT algorithm is given in Fig. 1.

4 Performance Study

In this section, we present the performance study for SARMUT and the extension of classical algorithms to find association rules from sampled transactions-FAST. The

standard association rules mining algorithm we used is UApriori proposed in paper [6]. The experiments were conducted on a machine with 3.10 GHz CPU and 4G main memory installed. The operating system is GNU/Linux.

4.1 Experimental Methodology

The final sampling ratios chosen are 2.5%, 5%, 7.5%, 10% and 10%. Preliminary experiments showed that a value of k=20 for the group size worked well in both FAST and SARMUT, and we therefore use this value throughout. We select 30% transactions of the whole dataset to obtain the initial sample for SARMUT and FAST.

Two data sets were used in the experiments-T40I10D100K and T25I15D320k, which were modified in literature [6]. For each dataset and algorithm, we run the algorithm 50 times, and then for each result we adopt the averages of the 50 observations.

Notation 2. Denoted by *S_T* the number of frequent itemsets that sampling-based mining algorithm produced from sample *S* and *D_T* the number of frequent itemsets that standard mining algorithm produced from the whole dataset. And let *T_T* be the number of the association rules both found by the two kinds of algorithms. We take the measure of accuracy as follows:

$$\text{Accuracy 1} = 1 - \frac{|S\_T - T\_T| + |D\_T - T\_T|}{S\_T + D\_T}, \tag{7}$$

4.2 Experimental Results

Accuracy vs. Sampling Ratio

Table 1 and 2 illustrate the accuracies of the different algorithms on the synthetic databases. As shown in the figures, SARMUT outperforms FAST in most cases. It also can be seen from the figures that as the final sampling radio increases, the algorithms get better accuracy, because large samples can reflect the whole database well and truly and they contain more useful information. The similar trend can be observed in variety of min*Sup* with respect to accuracy, for the higher min*Sup* lead much fewer patterns produced both by SARMUT and FAST with expected support ≥ min*Sup*.

Execution Time

Table 3 shows the sampling part execution and total execution time (sampling plus subset frequent itemsets mining) of SARMUT, FAST and UApriori on T40I10D100K database as a function of the final sampling ratio. We consider only 30% transactions of the dataset as the initiative sample in both SARMUT and FAST. Experimental results indicate that, when *final_size* decreases, more transactions should be removed from the initiative sample, and thus longer runtime for sampling part is required. Moreover, both SARMUT and FAST run much faster than the non-sampling algorithm UApriori.

Table 1. Accuracy 1 vs. Sampling Ratio on T40I10D100K

final_size	2.5%		5%		7.5%		10%	
min*Sup*(%)	SARMUT	FAST	SARMUT	FAST	SARMUT	FAST	SARMUT	FAST
0. 4	0.846	0.824	0.888	0.871	0.909	0.886	0.918	0.901
0. 6	0.857	0.836	0.895	0.888	0.915	0.903	0.927	0.919
0. 8	0.865	0.855	0.903	0.895	0.921	0.916	0.930	0.925
1.0	0.871	0.866	0.911	0.904	0.929	0.921	0.936	0.929
1.2	0.883	0.870	0.915	0.908	0.933	0.923	0.943	0.934
1.4	0.892	0.879	0.930	0.917	0.938	0.930	0.947	0.938
1.6	0.913	0.902	0.945	0.927	0.947	0.941	0.955	0.946
1.8	0.939	0.917	0.950	0.941	0.958	0.952	0.970	0.949
2.0	0.948	0.934	0.968	0.953	0.973	0.964	0.975	0.968

Table 2. Accuracy 1 vs. Sampling Ratio on T25I15D320K

final_size	2.5%		5%		7.5%		10%	
min*Sup*(%)	SARMUT	FAST	SARMUT	FAST	SARMUT	FAST	SARMUT	FAST
0.1	0.779	0.715	0.858	0.824	0.884	0.869	0.905	0.888
0.2	0.864	0.859	0.905	0.900	0.923	0.920	0.933	0.931
0.3	0.864	0.859	0.905	0.901	0.924	0.918	0.934	0.930
0.4	0.867	0.857	0.908	0.902	0.927	0.919	0.935	0.930
0.5	0.874	0.864	0.912	0.903	0.927	0.922	0.939	0.930
0.6	0.888	0.878	0.924	0.913	0.936	0.932	0.949	0.938
0.7	0.904	0.891	0.929	0.924	0.949	0.939	0.951	0.946
0.8	0.917	0.903	0.945	0.934	0.952	0.946	0.960	0.954
0.9	0.944	0.926	0.967	0.947	0.968	0.956	0.970	0.964

Table 3. Execution of FAST,SARMUT and UApriori when min*Sup*=0.2% with T40I10D100K

final_size	FAST (sampling part)	SARMUT (sampling part)	UApriori on subset	FAST (total)	SARMUT (total)	UApriori
2.50%	30.02	40.3	20.28	50.3	60.58	351
5.00%	27.55	36.91	28.71	56.26	65.62	351
7.50%	25.07	33.51	36.55	62.07	70.51	351
10.00%	22.61	30.10	45.82	68.43	75.92	351

5 Conclusions

In this paper we studied the problem of finding association rule from existential uncertain data. We solved this problem with the sampling method and introduced SARMUT, which is a variant of FAST algorithm. Based on the two-phase trimming strategy of FAST, we modify the algorithm to fit for uncertainty and add the vertical and horizontal characteristics of the uncertain database to the “distance” function in order to remove “outlier”. The experimental results show that the sample we get is more similar to the whole database and contains more useful information so that the accuracy is better than the natural extension of FAST. Moreover, SARMUT runs much faster than the standard mining algorithm while its accuracy is still very high.

References

1. Chui, C.-K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 47–58. Springer, Heidelberg (2007)
2. Dai, X., Yiu, M.L., Mamoulis, N., Tao, Y., Vaitis, M.: Probabilistic spatial queries on existentially uncertain data. In: Bauzer Medeiros, C., Egenhofer, M.J., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 400–417. Springer, Heidelberg (2005)
3. Leung, C.K.-S., Mateo, M.A.F., Brajczuk, D.A.: A tree-based approach for frequent pattern mining from uncertain data. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 653–661. Springer, Heidelberg (2008)
4. Leung, C.K.-S., Carmichael, C.L., Hao, B.: Efficient mining of frequent patterns from uncertain data. In: Proc. IEEE ICDM Workshops, pp. 489–494 (2007)
5. Chui, C.-K., Kao, B.: A decremental approach for mining frequent itemsets from uncertain data. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 64–75. Springer, Heidelberg (2008)
6. Aggarwal, C.C., Li, Y., Wang, J.: Frequent pattern mining with uncertain data. In: Proc. KDD, pp. 29–37 (2009)
7. Brönnimann, H., Chen, B., Dash, M., Haas, P., Qiao, Y., Scheuerman, P.: Efficient data-reduction methods for online association rule discovery. In: Data Mining: Next Generation Challenges and Future Directions, pp. 190–208. AAAI Press, Menlo Park
8. Chen, B., Haas, P., Scheuermann, P.: A new two-phase sampling based algorithm for discovering association rule. In: Proc. 8th ACM SIGKDD, Edmonton, Alberta, Canada, July 23-26 (2002)
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. VLDB, pp. 487–499 (1994)
10. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. ACM SIGMOD, pp. 1–12 (2000)
11. Leung, C.K.-S., Brajczuk, D.A.: Efficient algorithms for mining constrained frequent patterns from uncertain data. In: Proc., pp. 9–18 (2009)
12. Zaki, M.J., Parthasarathy, S., Lin, W., Ogihara, M.: Evaluation of sampling for data mining of association rules. In: Proc. 3rd KDD, pp. 283–286 (1997)
13. Toivonen, H.: Sampling large databases for association rules. In: Proc. 22nd VLDB, pp. 134–145 (1996)

Multi-agent System Collaboration Based on the Relation-Web Model

Maoguang Wang, Hong Mei, Wenpin Jiao, Junjing Jie, and Tingxun Shi

Key Laboratory of High Confidence Software Technologies of Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University,
Beijing 100871, China
wangmg@sei.pku.edu.cn

Abstract. In a multi-agent system, the agents have the capability to work collaboratively to solve the complex task. In recent years social computing provides a new perspective for multi-agent collaboration. This paper first introduces the role model and role relationships including role inheritance, role preference, role binding etc. Then a relation-web model is proposed referring to the social computing research. To a great extent, the relation-web model is used to simulate the social collaboration. The relation weight called trust degree is updated according to their collaboration result. When a complex task is assigned to some agents, the agents will construct the relation-webs for the sub-tasks completion. Finally, to test the relation-web model, an experiment is designed to predict the electricity consumption. The result proves the model to be available and useful while simulating the multi-agent collaboration process for solving the practical problem.

Keywords: relation-web model, multi-agent system, collaboration.

1 Introduction

One main challenge in multi-agent system is to build an effective model for multi-agent system collaboration. Multi-agent system collaboration has become an interdisciplinary research field including economic model, decision theory, logical reasoning and social intelligence etc. Now social computing, simulating complex social processes, provides a new perspective for multi-agent system research. The literature[1] pointed that agent-based social modeling focuses on the cognitive modeling of social behavior and individual agents' interactions. Social computing will strongly influence system and software developments

In multi-agent system, a lot of agents construct an agent society which has the capability to support the complex social behaviors simulation. The multi-agent system collaboration should make full use of the social computing research results. Our major challenge for multi-agent system collaboration is to design models and test the simulation process. Thus our work focuses on the multi-agent system modeling and testing the collaboration model using social computing technologies.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 illustrates the relation-web model in detail. Section 4 gives an

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 132–144, 2010.

experiment for electricity consumption prediction based on the relation-web model. Finally we conclude in Section 5.

2 Related Work

Considering different agent models and application scenarios, researchers have proposed various collaboration mechanisms [2], such as decision theory, logical reasoning and dynamic planning etc. From the social computing point of view, by integrating the social cooperation mechanism and multi-agent evolution for numerical optimization, Pan et al. [3] provided a social cooperation mechanism which imports the trust degree to denote the historical information for agents. In their work, the acquaintance net model was imported to construct and update the local environment of the agent. It improved the convergence rate by the cooperation characteristic of agents. Tao et al. [4] proposed an extended contract net protocol by introducing the acquaintance coalition policy, the trust degree parameter and the adjustment rules. Virginia Dignum and Frank Dignum [5] highlighted that social relationships determine different types of power links between roles and investigate what is the exact nature of this relationship between roles in an organization and what are the consequences of different structure forms.

Z.E. Chang and S.Li [6] developed a social collaboration model for group decision making process. Multi-agents could collaboratively work together to carry out a common task under a real-time 3D GIS environment. Fei-Yue Wang and Daniel Zeng et al. [1] have done some beneficial research in agent-based social modeling. They thought that the characterization of social structure and relations are typically represented via nodes and ties in network representation, such as complex social networks. Especially simulating complex social processes raises many research issues such as model specification, experimental design, and testing the simulation model. Now there exist the research gaps between individual cognitive modeling and multi-agent social simulation.

Although there is a lot of work on the combination multi-agent theory with social computing, these mentioned works are very useful and inspired us to do the research on the relation-web model to simulate the social network and support multi-agent system collaboration.

3 Relation-Web Model Construction

In this section we first define the basic concepts including role model and role relationships in support of building relation-web model. Then we introduce the relation-web model construction and its application in the multi-agent system collaboration.

3.1 Role Model

Roles are abstract entities which have some goals, responsibilities and capabilities etc. A number of agents are able to play a specified role. Likewise, one agent can play many roles in different scenarios.

Role: R::=<ID, Description, Goals, Responsibilities, Relationship, Collaborators, Authorities, Policies, Priority >

ID is a unique identifier;

Description gives an overview of the role;

Goals is one or several goals for the role to achieve;

Responsibilities describes the specific execution behaviors;

Relationship depicts the relations between the roles, such as role inheritance, role evolution and role conflict etc., which is helpful for role management and reasoning.

Collaborators represents the partners of roles;

Authorities shows the role's behavior permissions and prohibitions;

P**olices** can be regarded as some constrained rules, which includes action policy, goal policy and utility policy etc.;

Priority is used to constrain the role's preference.

Role Relationships:

Role inheritance: $\forall r_1, r_2 \in R$, if $r_1 \subseteq r_2$, the corresponding attributes r_2 is r_1's subset, r_2 inherits r_1 denoted as $r_1 \rightarrow r_2$. Role inheritance has the transitivity, and it provides ways of extending and classifying the existing roles. The users can define different sub-roles by inheritance.

Role assignment: $\forall r_1, r_2, r_3 \in R$, if the role r_1 assigns role r_3 to the agent playing role r_2, it is denoted as $r_1 \xrightarrow{r_3} r_2$. Role assignment provides flexible ways to manage the roles dynamically.

Role conflict: $\forall r_1, r_2 \in R$, if $\exists g_1 \in goals_{r_1}, g_2 \in goals_{r_2}$, and $g_1 \perp g_2$, we call r_1 and r_2 conflict denoted as $r_1 \perp r_2$, where $g_1 \perp g_2$ means r_1 and r_2 have conflict goals.

Role preference: $\forall r_1, r_2 \in R$, if $priority_{r_1} > priority_{r_2}$, it is said that r_1 prevails r_2 denoted as $r_1 \succ r_2$, where $priority_{r_1} > priority_{r_2}$ means r_1 has the higher priority than r_2 .

Role evolution: $\forall r_1, r_2 \in R$,if $r_1 \rightarrow r_2$, and r_2 extend r_1's goals, responsibilities, authorities and so on, we define r_1 evolves to r_2 denoted as $r_1 \xrightarrow{+} r_2$. Role evolution shows that the roles may be extended under specific conditions.

The agent is regarded as the autonomous entity which plays the specified roles in the system. Through role binding, the agent instances can be generated from roles.

Role-Agent binding: it is a map from the agents to the roles, the map function is denoted as bind(R): $R \rightarrow 2^{AE}$, 2^{AE} is the power set of the agent AE, that is to say, a number of agents can undertake one role. The role can dynamically bind to agents. Given that the set of agents As=$\{a_1, a_2, \ldots, a_n\}$ and the set of roles Rs = $\{ r_1, r_2, \ldots, r_m \}$,

the role binding matrix RB is defined as follows: $RB=Rs^T \times As = \{r_1, r_2, \ldots, r_m\}^T \times \{a_1, a_2, \ldots, a_n\} = \begin{bmatrix} r_1a_1 & r_1a_2 & \ldots & r_1a_n \\ r_2a_1 & r_2a_2 & \ldots & r_2a_n \\ & \ldots & & \\ r_ma_1 & r_ma_2 & \ldots & r_ma_n \end{bmatrix}$,

$$r_i\, a_j = \begin{cases} 1 & \mathrm{r_i\ binds\ to\ a_j} \\ 0 & else \end{cases} \quad (1 \le i \le m, 1 \le j \le n \quad m,n \in \mathbb{N})$$

In this definition, r_ia_j represents whether r_i binds to a_j or not. If ra=1, agent a inherits the r's attributes. For example, the r's goals is mapped to a's goals. That is, $\forall\ g_r \in goals_r,\ \exists g_a \in goals_a \Rightarrow bind(g_r) \rightarrow g_a$.

For example, in an international conference there are various roles such as conference chair, session chair, author, presenter and so on. When a paper is accepted and the author registers for the conference, the author's role evolves to a paper presenter. The role of presenter can bind to specific students or professors. The relationship between roles is depicted as figure 1.

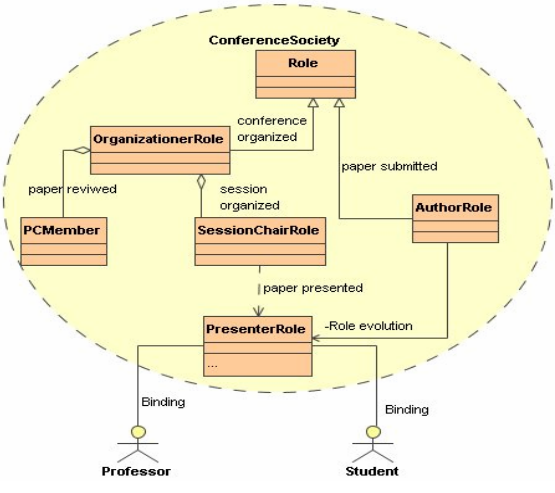


Fig. 1. Role relationships in the conference

In this figure we can see the concrete professor and student agents bind with the roles dynamically in the conference scenario. Furthermore they will play different roles while their attributes change.

3.2 Relation-Web Model

Drawing on the experience of social computing and social intelligence research result, we build dynamic relation-web model to support multi-agent collaboration. The relation-web model describes the related agents to work collaboratively in figure 2.

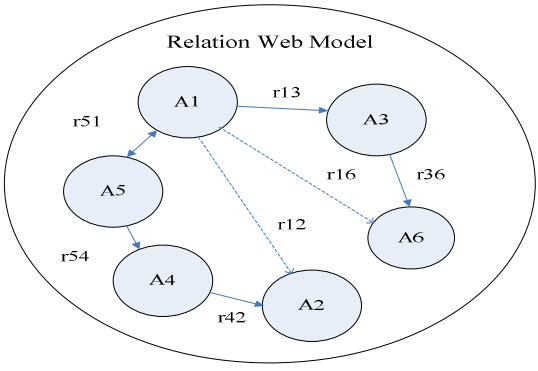


Fig. 2. Relation-web model

Relation-web model: RG= (As, Vs) , where As={A_1, A_2, …, A_n}, A_i is the agent, i=1,2,…,n; Vs={V_1, V_2, …,V_n}, V_i is the set of the agent A_i's relation weights between A_i and its acquaintances, V_i={r_{ij}, i, j=1, 2,…,L, j≠i, L is the number of A_i' s acquaintances}. The value of r_{ij} shows the relation intensity and interaction properties such as competition or coordination. Different from the proposed similar model[7],the relation-web model has reflexivity, non-symmetry and transitivity.

In fact an abstract relation-web model can be constructed by abstract roles. Furthermore, the concrete models can derive from the abstract model by role binding. However, to illustrate the model's essential characteristics, we just use the agents to construct the relation-web model directly. In Figure 2 A_1 establishes the relationship with A_3 and A_5 directly, A_4 establishes a relationship with A_2 directly. The dashed lines r_{12}, r_{16} represent that A_1 can interact with A_2 , A_6 by the relation transitivity, r_{16}= $r_{13}\odot r_{36}$, $\odot$ is a composite operator. According to the agents' interactions[8], r(A_i, A_j) also called trust degree is defined as r(A_i, A_j)= $\begin{cases}[-1,0\rangle & Competition \\ 0 & NA \\ (0,1] & Coordination\end{cases}$ .r∈ $[-1,0)$ means that the agents will compete to achieve the specified goal. If r∈ $(0,1]$, it means that the agents will collaborate to achieve the common goal. Otherwise if r=0, it shows that the agents have no relations at the moment. For instance, given a football match the relation weight set {-1, -0.5, 0, 0.5, 1} is used to represent the {intense competition, competition, independent, coordination, full collaboration}.

Supposed that agent is simplified as A_i=<AID_i, Cap_i, $Addr_i$, $Rw_{\underline{i}}$>, the parameters represent the agent identifier, capability, address and the set of relation weights respectively. Cap_i={C_1,C_2,…,C_m} will judge whether the agent has the capability to fulfill the given task. $Rw_{\underline{i}}$={Acq_i, R_i}, the agent A_i's acquaintances Acq_i and the set of relation weight R_i={r_{ij}, j =1, 2, …, |$Acq(A_i)$|, j≠i}. If $r_{ij}\geq 0$, the higher trust degree means that the agents have more probability to work collaboratively. When the agents need to collaborate to fulfill the assigned task, it will search its acquaintance set. Given that ι =<A_0, A_1, …, A_n>is the relation chain from A_0 to A_n, $L(\iota)$=|ι| is the

length of $L(\iota)$. The minimum distance from A_i to A_j denoted as $D(A_i, A_j) = \min(L(\iota))$ [8]. If $D(A_i, A_j) \neq \infty$, it is said that A_i can reach A_j.

Autonomic society[6] $AS=<AEs, Rel, DF>$, AEs is the finite set of agents, Rel is the set of trust degree, DF is a special agent responsible for agent registration, index and lifecycle management. When agents work collaboratively to fulfill the complex task, it will increase the communication cost during task decomposition and execution. If $|Acq(A)| \in [1, |AEs|-1]$, the average execution cost(task) $\in [O(1), O(|AEs|/2)]$. Therefore, we should make a trade-off between the space complexity and the cost of maintaining the acquaintances.

In practical applications, considering the trust degree changes dynamically, we should adjust the relation-web model dynamically to meet the trust degrees changes. For complex multi-agent systems, agents usually have no global knowledge and trust degrees will not be obtained accurately beforehand. When agents work together, it will succeed or fail. According to the final result, we will increase the trust degree if they succeed finally. Otherwise the trust degree will decrease. This will insure that the agents with higher success rate will increase the relation weight gradually.

Let the trust degree of A_j at t moment is $T(A_i, A_j, t) \in [-1,1]$, and at the initial moment $t=0$, the trust degree of $A_i \rightarrow A_j$, $T(A_i, A_j, 0) = 0$. While agents work collaboratively, T will increase a value δ(reward) if they succeed. Otherwise T will decrease a value γ(penalty) . How to determine the values of δ and γ is a key problem.

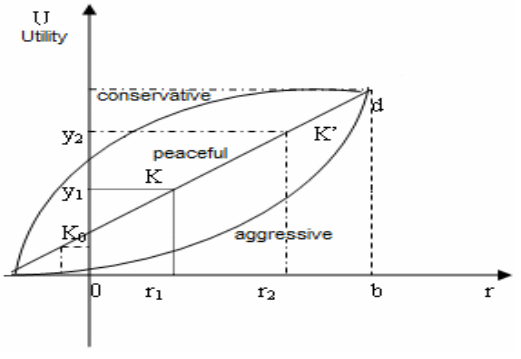


Fig. 3. The three types of agent

We divide the agents into three kinds: conservative agent who is allergic to the loss, peaceful agent whose utility is proportional to the adding profit, and aggressive agent that is more allergic to the gain profit. According to the kinds of agent, $\mathrm{val}(\delta, \gamma)$ is defined as follows.

$$\mathrm{val}(\delta,\gamma) = \begin{cases} \gamma > \delta, \text{Conservative agent} \\ \gamma = \delta, \text{Peaceful agent} \\ \gamma < \delta, \text{Aggressive agent} \end{cases}$$

When A_i and A_j collaborate successfully, $T(A_i, A_j, t+1) = \min(1, T(A_i, A_j, t) + \delta)$. While agent A_i and A_j fail to finish the task except other emergencies, $T(A_i, A_j, t+1) = \max(-1, T(A_i, A_j, t) - \gamma)$.

Furthermore, the trust degree can be defined as a relation weight function denoted as $F_t(x_1, x_2, \ldots, x_k)$, where $x_1, x_2, \ldots, x_k$ is the influence factors, which is capable of describing more complex relation-web models.

3.3 Relation-Web Model Based Collaboration

For the multi-agent system, the capability of solving complex problems is usually established by the agents' collaboration. In our work, once the task T_j is allocated, first we build agent society according to the relation-web model. The irrelevant agents will be excluded or considered with a low probability. While building the agent society, we mainly match the agents' capabilities and compare their trust degrees.

Let AS_T represents the agent society for task T, T is assigned to some agent $A_i \in AS_T$. S_i represents that the relation-web of A_i ($AS_T \supseteq S_i$). The agent$\in AS_T$ has some capability to support the task T completion. Once the relation-web is built, the task is decomposed into some sub-tasks according to the allocation policy. And the agents will be responsible for completing the specified sub-tasks. Finally all the returned sub-results will be summarized. Figure 4 illustrates multi-agent system collaboration for task T based on relation-web model.

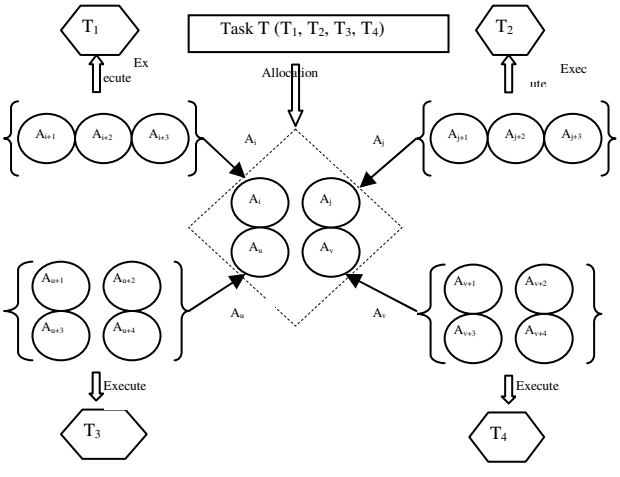


Fig. 4. Multi-agent system collaboration

The execution steps are as follows:

(1) The task T is decomposed into four sub tasks T_1, T_2, T_3 and T_4.

(2) DF agent assigns tasks to the agent society including four agents A_i, A_j, A_u and A_v.

(3) According to the respective relation chain, A_i, A_j, A_u and A_v search and match the capability of the acquaintances to finish the sub-tasks.

Table 1. Basic roles in the relation-web model

ID	**PSA(Power Company)**
Description	According to the power consumption, the PSA role suggests the power price to the government
Goals	Suggest the reasonable power price
Relationship	CA←, →PMA
Collaborators	CA, PMA
Responsibilities	PSA=(WaitStartStep·QueryCustomer·GetCustomerConsumption·CalculateStepTotalConsumption·SubmitTotalComsumption·SendPriceAdvice)$^{+}$
Authorities	Read: PowerPrice PersonalConsumption Write: StepTotalConsumption Send: PriceAdvice
Polices	PriceMeetMarket==TRUE
Priority	Low
ID	**CA(Electicity Consumer)**
Description	According to the personal income, power price and social influence, the CA role computes the power consumption.
Goals	Evaluate the power consumption
Relationship	→PSA, →PMA
Collaborator	PSA, PMA
Responsibilities	CA=(WaitStartStep·ReceivePrice·ContactNeighbour*·ConsumePower·ReplyNeighbour·SendPersonalComsumption)$^{+}$
Authorities	Read：PowerPrice NeighboursList PersonalParameters LastStepConsumption StepID Write：StepPersonalConsumption SocialWeights
Policies	PersonalPowerConsumption>0
Priority	Medium
ID	**PMA(Government Sector as the price maker)**
Description	According to the constumer power consumption and the suggested power price, the PMA role makes the power price.
Goals	Specify the power price
Relationship	CA←,PSA←
Collaborator	CA,PSA
Responsibilities	PMA=(ContactCustomer*·GetCustomerConsumption·ListenPriceAdvice·SpecifyPowerPrice)$^{+}$
Authorities	Read: PowerConsumption CompanyPrice Write: SpecifiedPrice
Policies	PowerPriceAcceted==TRUE
Priority	High

(4) Match capability, sort and select the acquaintances with higher trust degree. A_i, A_j, A_u, A_v will form their respective relation-webs, which will be responsible for T_1, T_2, T_3 and T_4.

(5) The agents collaborate to finish the sub-tasks.

(6) After they all fulfill the sub-tasks, the result are summarized and evaluated.

If the utility of A_i, A_j, A_u and A_v for executing sub-tasks is cost($T_{i,j,u,v}$) respectively, the utility of task T is Cost(T)= $\frac{\sum_i \text{cost}(T_i)}{n}$, n=4.

4 Experiment Design and Analysis

To test the relation-web model in a real application, we design the experiment for electricity forecasting. Firstly, we analyze and define the roles including PSA, CA, PMA in power supply chain. The basic roles depicted in table 1.

PMA charges for making the electricity policy between CA and PSA. PSA provides power to CA, and CA evaluates the power demands according to the family income, electricity price, household appliances, etc. PMA takes a variety of effective measures to regulate the electricity price. The agents build a relation-web model and work collaboratively while playing different roles. Figure 5 depicts the collaboration scenario.

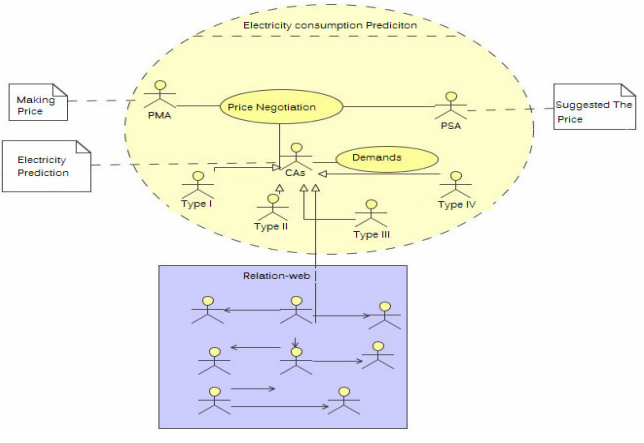


Fig. 5. The relation-web model for electricity consumption prediction

The electricity consumption is affected by a lot of factors. Usually, the electricity consumption can be expressed as C=f(P, Z), where C is the quantity of electricity consumption, P is the electricity price, Z represents the other influence factors including income level, household appliances, habits of using electricity and so on. The electricity demand of CA is expressed by the formula: $C(i,t) = a + bP(i,t) + cZ(i) + \varepsilon(i,t)$, $C(i,t)$ is the family i's electricity

consumption in time interval t, $P(i,t)$ is electricity price, $Z(i)$ is the user attribute eigenvector, a, b, c are constant coefficients, $\varepsilon(i,t)$ is the error term.

The literature [9] builds measurable short-term economic model from user requirement, which considers the influences including income and electricity price. They used the model to predict the electricity consumption without considering other social influence factors.

In fact, every family's electricity consumption is not only affected by the consumer's income and electricity price, but also affected by the number of family members and household appliances, saving habits and education etc. By analyzing the influencing social factors, the calculation model of CA's electricity consumption is represented as $C(i,t) = a + bP(i,t) + cZ(i) + dS(i,t) + \varepsilon(i,t)$, where $S(i,t)$ is social factor vector, d is coefficient, the others are same as the model mentioned above. The influencing factors include family income, policy and other social factors such as education, advertisment etc.

We build the multi-agent relation-web model experiment for electricity predition: PSA, CA and PMA represent the power company, the consumer, the government sector respectively. SA is used to integrate the final result. SA sets the related parameters for every participant and controls the simulation process. PMA makes the electricity price. CA estimates electricity requirement accoring to the relation-web model. CAs sends individual electricity demand to SA, SA calculates the total electricity consumption. Figure 6 depicts the structure of multi-agent relation-web platform.

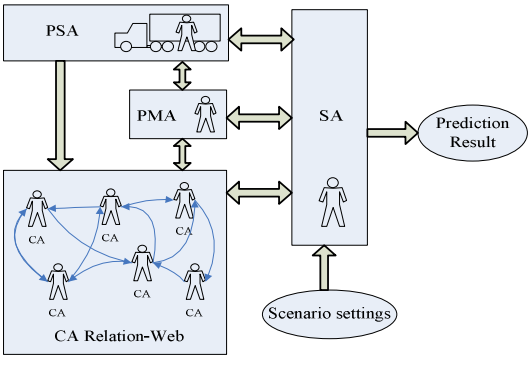


Fig. 6. Relation-web experiment platform

MAGE [10] is a multi-agent platform that is used to develop the relation-web model. CA represents every family and they influence each other. The dynamic influence is quantified by changing relation weight. In this experiment there are two kinds of influences on electricity consumption, one is direct influence from the government policy and education, the other is indirect influence from the neighbours.

In the experiment, there are four types of consumers, the diagram distinguishes these four types with different colours. Figure 7 depicts the relation-web construction.

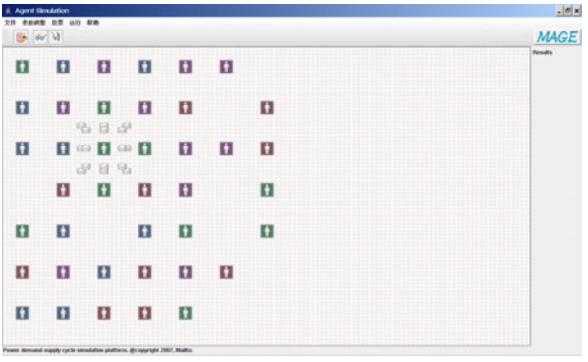


Fig. 7. Relation-web for multi-agent collaboration

As figure 7 depicted, each CA fixed on a grid defined as CA(x, y), (x, y) is the coordinate of CA. To simplify the relation-web model, we define eight grids circling around CA(x,y) as the CA's neighbors. Thus, the maximum number of the neighbors is eight. Every CA is only influenced by its neighbors. For example, $CA(1,1)$ has three neighbors including $CA(1,2)$, $CA(2,1)$ and $CA(2,2)$. The influence is calculated by the trust degree. Every neighbor calculates its influence over $CA(1,1)$ and transfers the total value to $CA(1,1)$ which adds all the values as the influence from the neighbors. The society influence is denoted as $S(i,t) = f(\sum_{j=0}^{n} sw_j)$, where sw_j represents the trust degree for agent j to i, n is the total number of agent i's neighbors. According to the domain knowledge, we define the trust degree function as $f(x) = \frac{1}{1+e^{(-(x-5))}}, x \in [0,9]$.

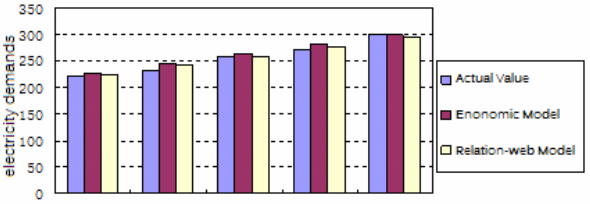
In the simulation experiment, each agent represents different customers which are bound to different roles. We use the model to predict the electricity consumption. The execution steps are as follows:

1) Initialize and construct the relation-web model. The initialization includes setting the parameters for relation-web scale, the types of agent playing different roles, the maximum response time and iteration steps and so on. Then generating the concrete agents randomly.
2) PMA makes the suggested electricity price and informs CAs and PSA.
3) Each CA gets the price, collaborates with the neighbours and calculates the society influences.
4) Each CA takes into account the social influence, estimates its own electricity demands and reports the demands to PSA.
5) PSA collects and calculates all of the consumers' electricity demands, inform the total demand to PMA, and give some suggestions if necessary.

6) PMA decides whether it should adjust the electricity price according to the feedback demands. Goto step 3), start next iteration.
7) Finish the simulation process and output the result.

During the collaboration, the consumers interact with its neighbours to get social influence value.

The experiment result gives the electricity demand prediction illustrated in figure 8.



Actual Value	22170	23280	25920	27220	30040
01-05with economic model	22845	24605	26488	28285	29937
01-05 with relation-web model(a=10,b=20,c=30, d=20)	22543	24163	25968	27710	29520

Fig. 8. Electricity prediction based on relation-web model

From the diagram, the electricity demand prediction based on the relation-web model has a better result which is closer to the real value.

The experiment result shows we can predict the electricity demands more accurately when considering the social influence factors. In fact when taking into account the social influences based on the relation-web model, the customer's habits will be calculated for the precise prediction.

5 Conclusions and Future Work

Multi-agent system collaboration has become an interdisciplinary research and application field. Referring to the social collaboration mechanism, we provide a relation-web model for simulating simple social network. And the experiment result gave a better prediction for real applications. To improve the efficiency, now we mainly adopt a top-down method during the task allocation in the experiment. However, an efficient relation-web model should be constructed not only by a top-down method, but also constructed and trained by the bottom-up method from the beginning. Our future work will focus on the dynamic relation-web model update by combing these two methods and will take advantage of the fuzzy rules and semantic reasoning to describe complex relation-web.

Acknowledgments. This work is supported by the National Basic Research Program of China (973) under Grant No. 2005CB321805, 2009CB320703, the Science Fund for Creative Research Groups of China under Grant No. 60821003, the High-Tech Research and Development Program of China under Grant No. 2009AA01Z139 and China Postdoctoral Science Foundation No. 20090460151.

References

1. Wang, F.-Y., Zeng, D., Carley, K.M., Mao, W.: Social Computing: From Social Informatics to Social Intelligence. IEEE Intelligent Systems, 79–83 (2007)
2. Wang, M., Shi, Z., et al.: Agent Grid Collaborative Schema Based on Policy. In: Communications in Computer and Information Science. 3rd International Conference on Intelligent Computing, vol. 2, pp. 469–478 (2007)
3. Pan, X.-Y., Jiao, L.-C.: Social cooperation based multi-agent evolutionary algorithm. Journal of Xidian University 36(2), 247–280 (2009)
4. Tao, H., Wang, Y., Guo, M., Wang, H.: A multi-Agent Negotiation Model Based on Acquaintance Coalition and Extended Contract Net Protocol. Journal of Computer Research and Development 43(7), 1155–1160 (2006)
5. Dignum, V., Dignum, F.: Coordinating tasks in agent organizations Or: Can we ask you to read this paper? In: Coordination, Organization, Institutions and Norms COIN@ECAI 2006 (2006)
6. Chang, Z.E., Li, S.: Social collaboration model support for collaborative 3D GIS environment based on multi-agent method. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, vol. XXXVII, Part B2, pp. 1059–1063 (2008)
7. Chen, G., Lu, R.: The Relation Web Model-An Organizational Approach to Agent Cooperation Based on Social Mechanism. Journal of Computer Research and Development 40(1), 107–113 (2003)
8. Wang, M.: Research on Modeling and Collaboration in Autonomic Computing. Phd Dissertation, Chinese Academy of Sciences (2007)
9. Song, Y., Gu, A., Wu, Z.: Urban residential electricity consumption forecast model. Electric Power 39(9), 66–70 (2006)
10. Shi, Z., Zhang, H., Dong, M.: MAGE: Multi-Agent Environment. In: ICCNMC 2003, pp. 181–188. IEEE CS Press, Los Alamitos (2003)

Research on A Novel Multi-Agents Dynamic Cooperation Method Based on Associated Intent

Weijin Jiang[1], Xiaoling Ding[1], Yuhui Xu[2], Wang Chen[2], and Wei Chen[2]

[1] Institute of computer application, Hunan University of Commerce, Changsha 410205, China
linxh_yyy@126.com
[2] School of Art Layout, Hunan University of Commerce, Changsha 410205, China

Abstract. With the information explosion speeds up the increasing of computing complexity rapidly, the traditional centralized computing patterns are under great pressure to process those large-scale distributed information. However, the agent-based computation and high-level interaction protocols foster the modern computation and distributed information processing successfully. The multi-agent system (MAS) plays an important role in the analysis of the humaninteraction theory and model building. This study focuses on the formal description of MAS, the conflict-resolving mechanisms and the negotiation in MAS. The communication between agents has some special requirements. One of them is asynchronous communication. Used communication sequence process (CSP) to descript a model of agents communication with shared buffer channel. The essence of this model is very suitable for the multiagents communication, so it is a base for our next step job. Based on the communication model, explored the distributed tasks dealing method among joint intention agents and with description of relation between tasks we give a figure of agents' organization. Agents communicate with each other in this kind of organization. The semantics of agent communication is another emphasis in this paper. With the detailed description of agents' communication process, given a general agent automated negotiation protocol based on speech act theory in MAS, then we use CSP to verify this protocol has properties of safety and liveness, so prove it is logic right. At last a frame of this protocol's realization was given.

Keywords: Multi agent system, Automated negotiation, CSP calculus, Liveness, Joint intention.

1 Introduction

The theory of Multi-Agent Automated Negotiation involves extensive applying fields and many kinds of methods. The theory mainly lies in Argument Based Automated Negotiation [1], Game Theoretic Models and Heuristic Approaches. In application, it can be divided into two categories [2], Agent's Negotiation within MAS and Self-interested between different MAS [3-5]. Those theories supporting the interior collaboration of MAS are like Self-interested, Joint Intentions and Shared Plans, no matter which are have differences, they have been working under the premise of

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 145–152, 2010.

identical intention and target of Agent within MAS [6]. This text will discuss the Joint Intentions in Multi-Agent Automated Negotiation of MAS [7].

If Multi-Agent in MAS interacts successfully, there must be three conditions demanded to be satisfied as below: Communication Structure, that is, how to dispatch and take over information between Agent[8]; Communication Language, that is, Agent is required to understand the signification of the information; Interaction rules, that is, how to organize the conversation between Agent [9].

Regarding to the research of Agent Communication Structure, we have proposed MAS communication model in the previous parts [10-12]. In the second section, it will be stressed to analyze Agent's asynchronous communication mechanism. As to the research of Agent Communication Language, presently there have been many abroad, like KQML, FIPA, ACL, Agent Talk, etc., so the language is not the emphasis in our text. Then, research of Interaction Rules is the second emphasis in the text. In the third part, the text will set forth the agreement of Agent Automated Negotiation and its validation. In the forth part, it illustrates and analyzes the complete frame of Agent Automated Negotiation. The fifth is the conclusion of the text.

2 Agent Communication Mechanism

2.1 Agent Communication State

Agent is a status course which can accomplish the task automatically with the ability and agreement of communication, for example, P_{A} represents the course of Agent A.

The course of Agent make the Agent's ability which can be marked as $Ability^{P_A}$ and $TASK^{P_A}$ means to be able to fulfill the task.

The moving status of the static Agent in MAS can be classified as Active, Wait and Run. Agent in the Wait status will be activated after receiving the requests from other Agent and then run. Agent in Run status will negotiate with other Agent or provide services according to the Try-best principle. $State_{outer}$ stands for the Run status of Agent:

$State_{outer}$::=*Wait* | *Active* | *Run*

Agent's collaborating course observed from the outer MAS is the process that Agent runs in the $\mathrm{I_{outer}}$=$\mathrm{Stae_{outer}}$*

In an Agent's collaborating process with Safety and Liveness, the circulation of *Wait*→*Active*→*Run*→*Wait* in $\mathrm{I_{outer}}$ will appear at least once to Agent's launch and acceptance.

Obviously, in the circulation of *Wait*→*Active*→*Run*→*Wait*, if any one part of Agent can not fulfill the circulation, it means something happened unexpectedly cause the deadlock or livelock to the system during the collaborating process, so the theorem attested.

Example 1 Agent A past passage C to transmit one thing dispense with responsive notify m to Agent B. Agent A starting tenor is P_A and agent B starting tenor is P_B. The wholly cooperating process for: $P_{\mathrm{A}}.(c) \parallel P_{\mathrm{B}}(c)$, with CSP:

$$P_A = Wait \xrightarrow{o} Active \longrightarrow Outer?\, x \longrightarrow c!\,(m) \longrightarrow Wait$$

$$P_B = Wait \xrightarrow{\bar{o}} Active \longrightarrow c?\,(m) \longrightarrow Wait$$

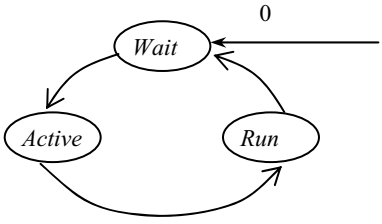


Fig. 1. Agent outer state transition fig. Sprung by events

In fig. 1, Outer stands for exterior entity relative to single Agent. o is the triggered event of outer, o and $\overline{o}$ is the coupling event, the said cooperating process possesses activity and security.

2.2 MAS Asynchronous Communication Mechanism

More and more application systems ask both corresponding sides of each other in a position to realize asynchronous communication mode. As a self-contained MAS communication structure, it is not only in a position to realize Agent's synchronous communication, but able to realize asynchronous communication. Miner'sðfiguring has realized transfer calculations by communication passage, which makes out that we can utilize Agent's asynchronous communication mode to realize synchronous. The asynchronous communication's ideal mode means that both corresponding sides own one infinite buffer queue. However, it is unpractical to deploy such infinite buffer queue to each Agent, whereas to share buffer channel may realize Agent's transfer between asynchronous communication and synchronous communication better.

Buffer channel C is such an Agent which set independent state switch and message buffer to all its relevant Agents and transmit messages for these Agents.

Example 2 In Example 1, utilizing a buffered passage C is to realize communication process. This example can realize the asynchronous communication between Agent A and Agent B, and the whole collaborating process is: P_A (C) $\| P_C \|$ P_B (C), as showed below:

Utilizing buffer channel may realize manifold asynchronous communication modes. Introductions of transmit message *m* through buffer channel as below

$P_A = Wait \xrightarrow{o} Active \rightarrow Outer?\ x \rightarrow C!\ (m) \rightarrow Wait$

$P_C = Wait \rightarrow C?\ (m) \rightarrow \text{if}(P_B.State_{outer} = Wait)\ then\ C!(m) \rightarrow Wait\ else\ Wait$

$P_B = Wait \xrightarrow{\overline{o}} Active \rightarrow C?\ (m) \cdot Wait$

The above process shows that Agent can realize asynchronous communication between Agents by use buffer passage. P_C stands for buffer channel tenor.

The synchronous communication between Agents asks Agents themselves shall be clear about each other's corresponding location. If a MAS system owning N (numerous) Agents would like to realize point-to-point communication between Agents, there will be N^2 channels needed to set up, of which so many will complicate

the realization of Agent extremely. Using shared buffer channel can be good for realizing channel's transmission between Agents.

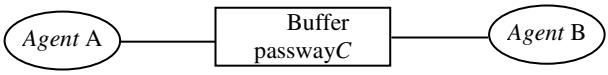


Fig. 2. Realize asynchronous communication between agents by using buffer passage

Example 3 Agent A transmits a piece of service request and its own corresponding location to its relative buffer channel, and then waits for despondence. If Agent C will be able to content Agent A with service request, Agent C shall utilize Agent A's corresponding location dispatch accepted message to Agent A. Therefore, there will be a synchronous communication channel set up between Agent A and Agent C. See Fig. 3 of communicating process.

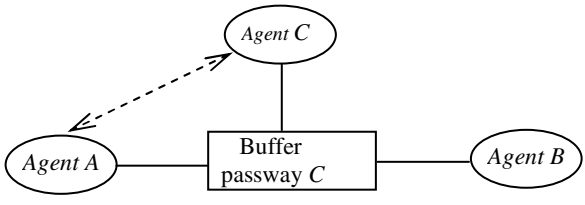


Fig. 3. Erect synchronous communication channel by utilizing buffer channel

3 MAS Interior Cooperation Mode

3.1 Agent Cooperating Principle

When Multi-Agent in MAS begins cooperation, for the reason that there is a conform joint intension between Agent, the process of Multi-Agent in MAS works according to the principal of "From each according to his ability, abide by the law and behave oneself", that is, each Agent is trying its best to cooperated with other Agent.

The cooperation between Agents is aimed at fulfilling a certain tasks. Because tasks can be divided into different but related sub tasks, the tasks from Agent's point of view can be described as following: a material task can be regarded as sub-tasks' assembling depending on different ability of Agent in MAS. Combining divided-task-oriented Agent in compliance with sub tasks will be in position to form a furcation tree of $k(k\geq 2)$. Relation between sub tasks is relation with or to time sequence. Agent's organizing relation is determined by the relation between tasks. Description of sub tasks as below:

(1) The sequential relationship of the tasks （<） , manifests that Agent B's task can not be begun before fulfilling Agent A's task. Formalization to be described below:

$TASK_{P_A} < TASK_{P_B} \models P_A; P_B$

Thereinto: TASK_{P_A} and TASK_{P_B} respectively means the start-up tenor P_A and P_B of Agent A and Agent B are used to fulfill tasks.

(2) The relation of "AND" between tasks （V）, indicates that Agent A and Agent B perform simultaneously sub task P_{A} and P_{B}. After completing the sub tasks, Agent C begins their common and subsequential task P_{C}. Formalization described as below:

$$\mathrm{TASK}_{P_A} \vee \mathrm{TASK}_{P_B} \models (P_A \parallel P_B) < \mathrm{TASK}_{P_C} \models (P_A \parallel P_B) < P_C$$

(3) The relation of "OR" between tasks （∧）, indicates that Agent A and Agent B with the relation of "OR" perform simultaneously sub task P_{A} and P_{B}, no matter which is fulfilled first, Agent C can begin its subsequential task P_{C}. Formalization described as below:

$$\mathrm{TASK}_{P_A} \wedge \mathrm{TASK}_{P_B} \models (P_A < \mathrm{TASK}_{P_C}) \parallel (P_B < \mathrm{TASK}_{P_C}) \models (P_A < P_C) \parallel (P_B < P_C)$$

Example 4 TASK_{MAS} means the task can be fulfilled by MAS, which is divided as the task tree seen in Fig. 4.

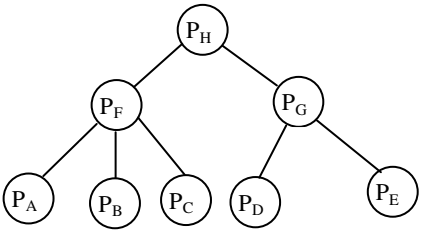


Fig. 4. Task Tree

(1) Relations between tasks: $(((P_A \vee P_B \vee P_C) < \mathrm{P_F}) \vee ((\mathrm{P_D} < P_G) \wedge (\mathrm{P_E} < \mathrm{P_G})))P_{\mathrm{H}}$

(2) With CSP describing $\mathrm{TASK}_{\mathrm{MAS}}$ as follows:

$TASK_{MAS}=(((P_A \parallel P_B \parallel P_C); P_F) \parallel (P_D; P_C \parallel P_E; P_G)); P_H$

From the above mentioned：MAS is a task processing distributive system. The Agent's ability can be realized by its corresponding tenor. The relations between tasks in MAS have determined that Agent is organized according to its dendriform communication topology which is the precondition for Agent's automatic negotiation.

3.2 Automatic Negotiation in Agent Protocol

Agent automatic negotiation is the main method for multi-Agent to negotiate, which focus on three aspects lieing in negotiation protocol, negotiation object and negotiation policy. Negotiation protocol and negotiation object act as the textual points, but the negotiation policy is clampinged how to look for in Agent each from of negotiation space best in order to reach consistence, concretion content visible literature cited.

Present hypotheses 1 to ensure negotiation agent could each other have partner faith in against due to MAS interior Agent according to Try-Best principle proceed synergic, furthermore MAS possess concurrent combine intent.

Negotiation Agent knows each other in negotiation policy.

Be on the negotiation with the result that decision agent toward internetwork communication negotiatory condition of Agent automatic negotiatory course mission due to specific assignment require different communication quality guarantee AND specific network insurance. Text take mission negotiation AND internetwork communication negotiation as agent automatism negotiation in process two phase.

Definition 4 MAS interior agent automatic negotiation course could include two phases. The first phase is based on multi-Agent automatic negotiation whose negotiation object includes task starting time, task ending time and the relation of the tasks; The second phase is the negotiation of Agent's communicating conditions whose negotiation object include corresponding security policy and network service quality(Q_OS).

According to the top analysis talks about with the correlative language behavior academic theories, we say the Agent automatic negotiation correspondnce in the procedure to state row word certain for: request, promise, refuse, advise, counter advise. In view of agreement presence overtime event and agent unsolicited message transmission, so increase overtime （timeout） status and inform （inform） state row word that. Communication protocol engine of the communication process state as follows of the agent:

$State_{inner}$ ::=*Started*| *Requested* | *Accepted* | *Refused* | *Promised* | *Informed* | *Advised* | *CAd-vised* | *Timeout* | *Stopped*

Agent automatic negotiation protocol can be divided into information transmission layer, buffer channel layer and Agent negotiation protocol layer from bottom to top, of which buffer channel layer C is one of the needed layers between Agents to realize asynchronous communication. If it will realize point-to-point synchronous communication between Agents, it can do communication directly through channel C. As to the description of Agent automatic negotiation, it mostly focus on Agent negotiation protocol layer, while for the other layers, it only describes their services and running environment in brief. In essence, the function of Agent negotiation protocol layer is the description of process.

The Agent A describe with Agent B whole negotiation procedural not formal as follows: Agent A first of all dispatch negotiation beg of Agent B received solicit aback, toward request message proceed analyses, could as per three strain scene dispose to：the first thing, in the event of Agent B receivability the solicit of Agent A, those Agent B to Agent A dispatch take send , else dispatch thumb advise， down through upon, the service request block mode， of the such negotiation scene as conventional C/S. the second thing, Provide some Agent B can provide serve of instruct, but because of the restrict of the resource of system can't very much the serve, so the Agent B can put forward to Agent A the serve promises, the Agent A handles Agent B the commitment of serve can proceed very much: Reject or accept. the third thing, The Agent B thinks after analyzing the Agent A request Agent A some items modification within request empress, can satisfy the Agent A request still, like this Agent B after proceeding Agent A some items within request to modification,

conduct and actions the suggestion sends out to the Agent A. Agent A for the suggestion of Agent B can operation proceeding as follows: Accept, reject and put forward the counterproposal. either that of toward Agent A counterproposal, Agent B receivability, reject or set own the other one proposal for.

3.3 The Verification of Agent Automatic Negotiation Protocol

Utilizing process algebra to carry out formalization of communication protocol not only can state logic structure and time sequence nature of the protocol narrowly, but also is favorable to verify the protocol. The nature of the protocol includes Liveness and Safety. In the protocol system of liveness, its process algebra expression must own the recursion characteristic from initial state to the passing. If protocol stops executing a certain event and is unable to go on, the system will be dead locked. If protocol executes some certain events circularly and infinitely but is unable to return to initial state, the system will be alive locked. The system without dead lock or alive lock will be safe.

Consider protocol JIAANP =| | | Pi =(P1 | | | P2 | | | … | | | Pn) =| | | (PS | | PR), because will not carry on communication directly between Pi and Pj (i≠j) , can think that they are separate , namely can store in and lock or live and lock. So we may prove that if there is dead lock or alive lock appeared between promoter process P_S and acceptor process P_R of Negotiation.

Considering three kinds of different conversation scene Q1 in agreement JIAANP, Q2 , and Q3, among them, the simple message is sent and received in the execution course of Q1 and Q2, not forming circulation in the state changes picture of the agreement, so they will not be formed and extremely locked or lived the lock. There is proposing and counter proposal circulation in the execution course of Q3, it carries out course and may be formed and locked and lived and locked very much, the complexity because transmit for the agreement overtime proves the difficulty brought for the agreement, so we suppose: The transmission of the network is reliable , the feedback between Agent is in time, namely, logic exactness of the agreement that the prerequisite without incident in overtime comes down to prove in the agreement.

4 Conclusions

This text provides a common and communication-based Agent cooperation mode by studying mutual behavior of Agent cooperation. The text also uses some effective format ways to depict automatic negotiation protocol of Agent process and verify the validity of the protocol's logic. Finally, the text makes an implementation frame for this agreement. While using blackboard mode to realize buffer channel in this implementation frame, it provides a deployed agreement stack extra and at last it presents performance analysis and expandable analysis. In addition, as to negotiation between Agent in MAS, because the advantage difference of Agent group negotiating with Agent which has a conform joint intension has great differences on negotiation principle and strategy, the self-interested Agent's negotiation agreement between MAS is our next work under research.

Acknowledgments. This work is supported by the Sci. and Technology Plan Project of China under of Hunan Grant No.2009GK2002, and the Technology Innovation Foundation under Science and Technology Ministry Grant No. 09C26214301947.

References

1. Jennings, N.R., Faratin, P., Lomuscio, A.R., et al.: Automated negotiation: prospects. Methods and challenges. In: Pacific Rim International Conference on Artificial Intelligence (2000)
2. Grosz, B., Sidner, C.: Plans for discourse. In: Cohen, P., Morgan, J., Pollack, M. (eds.) Intentions in Communication. Bradford Books, MIT Press (1990)
3. Rosado, I.J., Bernal-Agustin, J.L.: Genetic Algorithms in Multistage Distribution Network Planning. IEEE Trans. Power Systems 9(4), 1927–1933 (1994)
4. Kinny, D., Ljungberg, M., Rao, E., Tidhar, S.G., Werner, E.: Planned team activity. In: Castelfranchi, C., Werner, E. (eds.) MAAMAW 1992. LNCS, vol. 830. Springer, Heidelberg (1992)
5. Dillenbourg, J.A.S.: A Computational Approach to Distributed Cognition. European Journal of Psychology Education 7(4), 252–373 (1992)
6. Bredin, J., Kotz, D., Rus, D., Maheswaran, R.T., Imer, C., Basar, T.: Computational markets to regulate mobile-agent systems. Autonomous Agents and Multi-Agent Systems 6(3), 235–263 (2003)
7. Jiming, L., Han, J., Tang, Y.Y.: Multi-agent oriented constraint satisfaction. Artificial Intelligence 136(1), 101–144 (2002)
8. Sierra, C., Faratin, P., Jennings, N.R.: A service-oriented negotiation model between autonomous agents. In: Boman, M., Van de Velde, W. (eds.) MAAMAW 1997. LNCS, vol. 1237, pp. 17–35. Springer, Heidelberg (1997)
9. Hoare, C.R.A.: Communicating sequential processes. Prentice Hall International, New Jersey (1985)
10. Jiang, W., Zhang, L., Pu, W.: Research on Grid Resource Scheduling Algorithm Based on MAS Cooperative Bidding Game. Science in China F 52(8), 1302–1320 (2009)
11. Jiang, W.J.: Research on the Learning Algorithm of BP Neural Networks Embedded in Evolution Strategies. In: WCICA 2005, pp. 222–227 (2005)
12. Jiang, W.J., Wang, P.: Research on Distributed Solution and Correspond Consequence of Complex System Based on MAS. Journal of Computer Research and Development 43(9), 1615–1623 (2006)

Automobile Exhaust Gas Detection Based on Fuzzy Temperature Compensation System

Zhiyong Wang[1], Hao Ding[2], Fufei Hao[3], Zhaoxia Wang[*], Zhen Sun[1], and Shujin Li[1]

[1] Tianjin Key Laboratory of Film Electronic and Communication Device, Tianjin University of Technology, Tianjin 300999, P.R. China
[2] College of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin 300072, P.R. China
[3] Tianjin Public Security Bureau, Tianjin 300090, China
cyw@eyou.com, dinghaobm@tju.edu.cn, tingzhuhao4412@eyou.com, elewz@nus.edu.sg, chenyu0532@163.com, lsjking370@sina.com

Abstract. A temperature compensation scheme of detecting automobile exhaust gas based on fuzzy logic inference is presented in this paper. The principles of the infrared automobile exhaust gas analyzer and the influence of the environmental temperature on analyzer are discussed. A fuzzy inference system is designed to improve the measurement accuracy of the measurement equipment by reducing the measurement errors caused by environmental temperature. The case studies demonstrate the effectiveness of the proposed method. The fuzzy compensation scheme is promising as demonstrated by the simulation results in this paper.

Keywords: Automobile exhaust gas, Fuzzy inference, Temperature compensation.

1 Introduction

There are more and more automobiles in the world, and the annual output of the automobiles has exceeded 50 millions [1]. The automobile exhaust pollution is becoming more and more serious. The accurate measurement and proper treatment of exhaust gas of the automobile are very important. Exhaust gas of automobiles is a main source of pollution, and that pollution is difficult to be treated because analyzing and detecting exhaust gas of the automobile is a complex technology based on many subjects. Many countries invested a lot of man power and financial power to solve the automobile exhaust problems, but there is still no technique that can be accepted by all the users.

Recently, infrared gas analyzer are widely used to detect and measure the components of the gases, but the measurement accuracy are influenced by factors arising from testing conditions, such as environment temperature, which may lead to erroneous gas detection [2], [3].

[*] Corresponding author Zhaoxia Wang became an Associate Professor in Tianjin University of Technology from 2004. She is now with the Department of Electrical & Computer Engineering, National University of Singapore, 119260.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 153–160, 2010.

Computational intelligence, which is known to be different from the conventional computing, called the soft computing, mainly includes fuzzy logic [4], Genetic algorithms [5], neural networks [6], and the combined method of any formers, such as fuzzy neural networks [7], is being rapidly developed in recent years. Fuzzy logic is a branch of computational intelligence. It is a powerful tool for modeling human thinking and perception. This makes it easier to perform tasks that are already successfully performed by humans. Human experience can be quantized into rules of the fuzzy inference system.

Fuzzy set and fuzzy logic have became one of the emerging areas, which spread across various research fields, including control, pattern recognition, and other engineering problems. It has been applied in traffic control, network planning and fault diagnosis [8], [9], but only a limited number of publications on the application of fuzzy inference on detection of automobile exhaust gas have been reported [10], [11]. Our previous work employed fuzzy inference system into infrared exhaust gas analyzer in order to minimize the effect of the environment temperature and has demonstrated that fuzzy compensation scheme is better than the non- fuzzy method [11]. In this paper our previous works [11] are extended and the more reasonable membership functions and rules are selected to improve the accuracy of the measurements.

The designed fuzzy temperature compensation scheme is shown in Fig. 1.

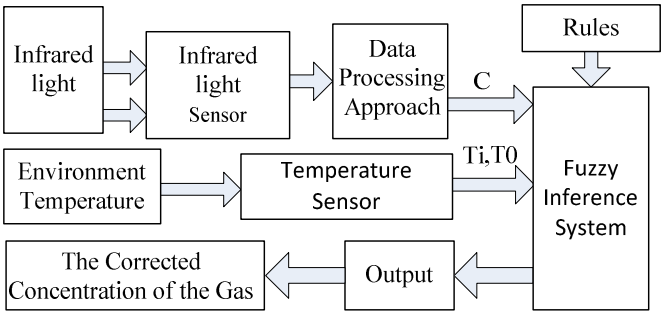


Fig. 1. Fuzzy temperature compensation scheme

There are two inputs and one output in the fuzzy temperature compensation scheme. One of the inputs is the density of the gas, which is obtained by infrared light sensor. The other input is the environment temperature, which is measured by using the temperature sensor. The rules are designed according to expert experience. According the output of the fuzzy inference system, the corrected concentration of the gas can be obtained.

2 Reviewing and Analyzing the Principle of Infrared Automobile Exhaust Gas Analyzer

The selective absorption of gas on infrared ray is used to analyze and process the data obtained from the sensor. The measured data is needed to be compensated because the error is too high when the variation of the temperature is high.

Inert gas does not absorb infrared energy, while the components of automobile exhaust gas are not only composed of inert gases, but also other different atom gases such as, as CO, CH and CO_2, etc. These gases can absorb the infrared energy in certain wavelength because they are not inert gases. The wavelength of the infrared energy, which can be absorbed, is called characteristic wavelength or feature wavelength. The absorption intensity is reflected by absorption coefficient. Due to the absorption of these gases, the energy of infrared ray will be reduced when the infrared ray go through the tested gases. The amount of reduced energy depends on gas concentration, the thickness of the gas chamber, and the absorption coefficient. The relationship is shown by Lambert-Beer's law [2], [3], [10], [11]:

$$\Delta E = E_0 - E = E_0(1 - C^{KL}) \tag{1}$$

where

E_0 : the initial infrared energy before going through the gases

E : is exit infrared energy after going through the gases

C: the gas concentration,

L : the thickness of the gas chamber

K: the absorption coefficient.

The absorption coefficient K is a very complicated parameter, which not only depends on the type of the gases, but is also related to the absorbed wavelength, and affected by the environmental temperature, pressure and other factors. Therefore, for the real working environment, in which both the temperature and pressure are varied, *K* directly affects the absorbed energy of infrared ray, ΔE. Huston pointed out that the absorption for the weak-lighted zone and the strong-lighted zone can be represented by the follow equation [10], [11]:

$$E = \int_{\Sigma} A(\lambda) d\lambda = cW^{1/2}(P + p)^q \tag{2}$$

$$E = \int_{\Sigma} A(\lambda) d\lambda = C + D\log W + Q\log(P + p)^q \tag{3}$$

where c , C , D , Q , q are constants determined by experiments. $A(\lambda)$ is the absorption function for certain wavelength. $d\lambda$ is infinitesimal increment of the wavelength. Σ is the wavelength range. P is the general environmental pressure. p is the measured gas pressure.

$$W = (pl/76) \times [273/(273 + t)] \tag{4}$$

where t is the environmental temperature; l is the thickness of the measured gas.

Equation (2) and equation (3) are employed for weak-lighted zone and the strong-lighted zone respectively [10], [11].

The above equations show that the environmental temperature is an important factor that influences the measurement accuracy of the sensor. The environmental temperature not only has an effect on the infrared ray absorption, but also directly affects the infrared radiation intensity and infrared detector accuracy.

The aim of this research work is to establish environmental temperature compensation scheme to eliminate the effect of environmental temperature.

When gas chamber is filled with pure nitrogen, the concentration of the testing gas is zero. For the fixed system parameter, gas absorption coefficient and gas chamber length L are fixed. So system parameter K can be obtained [10], [11].

Most of the equipment is operated at room temperature. In this paper, the room temperature 25℃ is selected as the standard temperature. The absolute error of measurement is:

$$e_i = C_{Ti} - C_{25^oC} \tag{5}$$

where C_{Ti} is the gas density measured at the temperature Ti, and C_{25^oC} is the gas density measured at the room temperature 25℃.

It has been demonstrated that higher the degree of the deviation from room temperature 25℃ for the environmental temperature, the higher the measurement errors are [10], [11]. Moreover, the denser the gas concentration is, the higher the measurement errors are. The compensation measurement must be considered because errors are too high to meet the demand of the measurement precision.

3 Temperature Compensation Scheme Based on Fuzzy Inference System

As mentioned in section II, for the different temperature, the sensors have different I/O characteristics. If the characteristic of C_{Ti} is ensured at working temperature T_i, we can get the corresponding output value C. The I/O characteristics can be ensured only under several limited number of temperatures. However, the I/O characteristics under any temperature T_i within the range of the work temperature can be estimated [11].

Fuzzy inference system [8]-[11] is summing up the experiences and quantizing them into a number of rules. After the membership function and rules are designed, the fuzzy inference system can complete the operation simply and rapidly. The outputs of the sensors can be corrected to requested precision by choosing a suitable fuzzy model and establishing appropriate rules.

The inputs and outputs of a fuzzy inference engine are linked by several inference rules, which are converted from expert knowledge in the following format [8]-[11]:

If Input is [(the environment temperature is very low) and (the concentration factor is very low)], then the output is [(The output of sensor is the function of this environment temperature and this concentration of gas))].

Absolute error [11] at different temperatures is used in this paper. The room temperature 25℃ is set to standard zero. The simulation data of the empirical errors at different temperature from -10 ℃ to 60℃ for different density of gas from 1.5% to 19.5% are shown in Table I [11]. The experiment data in Table I show that the error become higher with the gas density increasing.

The fuzzy temperature compensation scheme is shown in Fig. 1. Infrared light is measured by the infrared light sensors. Environment temperature is measured by the temperature sensor. Environment temperature and the density of the gases are inputs of the fuzzy inference system. The corrected concentration of the gases can be obtained by using the output of the fuzzy inference system.

According to the empirical data at different temperature, the two inputs of the fuzzy inference system, temperature and density of CO2 are represented by the fuzzy linguistic variables. The universe of discourse of Temperature from -10 ℃ to 60℃ are divided into six fuzzy sets NB, NS, ZO, PS, PB and PB2 as shown in Fig. 2.

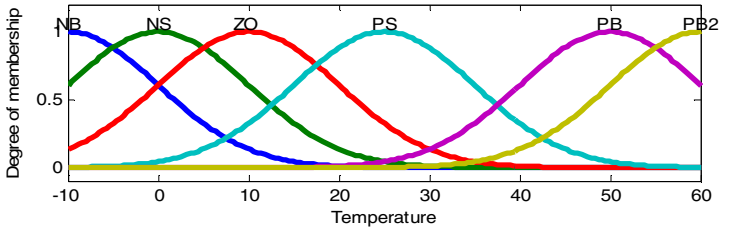


Fig. 2. Membership functions of the input: enviroment temperature

The universe of discourse of the concentration of gas from 0% to 20% are divided into four fuzzy sets S, M, B and VB as shown in Fig. 3. The gauss membership functions are used as shown in Fig. 2 and Fig.3. The universe of discourse of output is quantized into sixteen fuzzy sets and the membership function of the output is shown in Fig. 4.

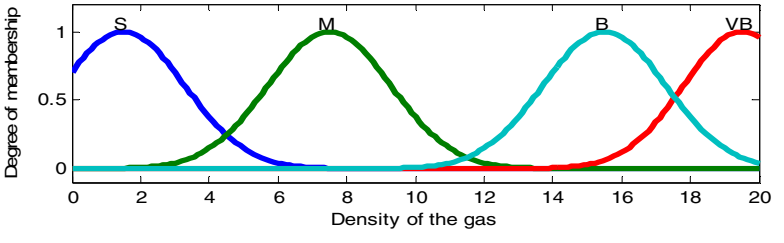


Fig. 3. Membership functions of the input: density of gas

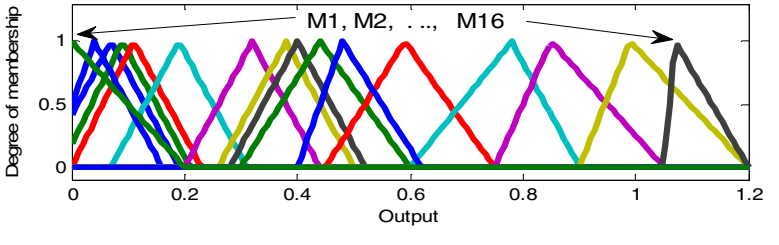


Fig. 4. Membership functions of the Output

The following sixteen rules are used in this fuzzy system:

If (input2 is PS) then (Output1 is M1)
If (input1 is S) and (input2 is NB) then (Output1 is M4)
If (input1 is S) and (input2 is NS) then (Output1 is M3)
If (input1 is S) and (input2 is ZO) then (Output1 is M2)
If (input1 is S) and (input2 is PB) then (Output1 is M5)
If (input1 is M) and (input2 is ZO) then (Output1 is M6)
If (input1 is M) and (input2 is NS) then (Output1 is M7)
If (input1 is M) and (input2 is NB) then (Output1 is M8)
If (input1 is M) and (input2 is PB) then (Output1 is M9)
If (input1 is B) and (input2 is ZO) then (Output1 is M10)
If (input1 is M) and (input2 is PB2) then (Output1 is M11)
If (input1 is B) and (input2 is NS) then (Output1 is M12)
If (input1 is B) and (input2 is NB) then (Output1 is M13)
If (input1 is B) and (input2 is PB) then (Output1 is M14)
If (input1 is VB) and (input2 is PB) then (Output1 is M15)
If (input1 is VB) and (input2 is PB2) then (Output1 is M16)

Fig. 5 shows the whole surface of the fuzzy inference system used in this temperature compensation method. The input 1 represents the temperature and the input 2 represents the gas density of CO2. Simulation experiment results demonstrate that using the limited empirical data in Table I, the errors at any temperature and for any density of the gas can be obtained by the fuzzy compensation system. In the temperature compensation scheme as shown in Fig.1, the function of fuzzy inference system is to estimate the errors causing by the temperatures.

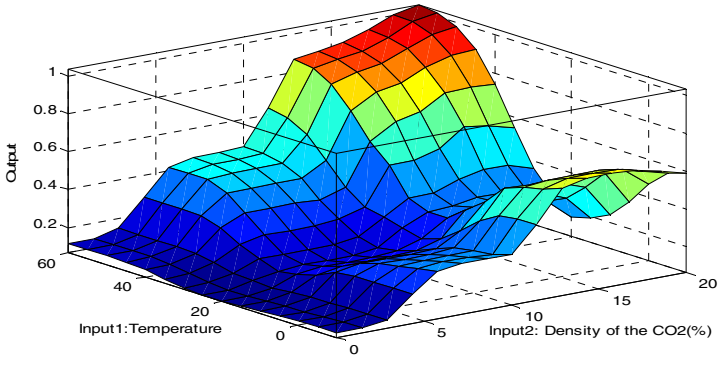


Fig. 5. Output surface of the Fuzzy Temperature Compensation System

According to the above results, the fuzzy temperature compensation scheme shown in Fig. 1 will be a feasible candidate for improving the measurement accuracy. Due to it simplicity, the scheme will be ideal selection for online application.

4 Conclusions

This paper discusses a fuzzy temperature compensation scheme, which can reduce the errors of the infrared automobile exhaust analyzer caused by environment temperature. All other automobile exhaust gas can be detected by this proposed method. Using the density of the CO2 as studying cases, the simulation results demonstrate the efficiency of this method. Using the limited empirical data, fuzzy system can improve the accuracy of the measurement. The universe of discourse should be divided into more fuzzy sets and more empirical data can be used to set up the fuzzy inference system for better result.

According to the Lambert-Beer law, it is obvious that the atmosphere pressure also can affect the measurement accuracy of infrared automobile exhaust analyzers. To consider the atmosphere pressure and to apply fuzzy temperature compensation system to the current automobile exhaust gas analyzer, further research works are needed.

Acknowledgment

This work is supported in part by the project of Tianjin Key Laboratory of Control Theory and Applications in Complicated Systems and Tianjin Key Laboratory of Film Electronic and Communication Device, Tianjin University of Technology.

References

1. Zhou, X., Bai, M., Zhang, Z., Deng, S., Huang, G.: Treatment of automotive exhaust gases by high-speed electrons. Journal of Electrostatics 57(3-4), 209–296 (2003)
2. Lin, Z., Yu, Y.: Development and Research of a New Type Non-dispersive Infrared CO2 Gas Concentration Analyzer. Sensor World 14(7), 13–14 (2008)
3. Zhao, R., Chu, J., Wang, R., Wu, C.: Purification of automotive exhaust by catalysis-assisted low temperature plasma. Gaodianya Jishu/High Voltage Engineering 33(2), 174–177 (2007) (in Chinese)
4. Wei, K., Wang, Z., Wang, Q.: Describing fuzzy sets using a new concept: fuzzify functor. Optoelectronics Letters 5(1), 69–72 (2009)
5. Wang, Z., Chen, Z., Yuan, Z.: QoS Routing optimization strategy using genetic algorithm in optical fiber communication networks. Journal of Computer Science and Technology 19(2), 293–297 (2004)
6. Wang, Z., Sun, Y., Wang, Z.-Y., et al.: The Research on Predicting Network Traffic Using Neural Networks. Journal of Opto-electronics. Laser 17(10), 1256–1258 (2006) (in Chinese)
7. Wang, Z., Sun, Y., Chen, Z., Yuan, Z.: The study of predicting network traffic using fuzzy neural networks. Journal on Communication 26(3), 136–140 (2005) (in Chinese)
8. Liu, X.-F., Ma, L., Mathew, J.: Machinery fault diagnosis based on fuzzy measure and fuzzy integral data fusion techniques. Mechanical Systems and Signal Processing 23(3), 690–700 (2009)

9. Zidani, F., Diallo, D., El Hachemi Benbouzid, M., Nait-Said, R.: A fuzzy-based approach for the diagnosis of fault modes in a voltage-fed PWM inverter induction motor drive. IEEE Transactions on Industrial Electronics 55(2), 586–593 (2008)
10. Shan, W., Jin, D.: Optimization method of fuzzy logic controller and its application in compensation of sensor. Transducer and Microsystem Technologies 26(6), 90–93 (2007) (in Chinese)
11. Li, S., Bai, M., Wang, Q., et al.: Detecing automotive exhaust gas based on Fuzzy inference system. In: 6th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2009), August 15-18, pp. 267–270 (2009)

Technical Efficiency Analysis of New Energy Listed Companies Based on DEA

Gao Chong[1,*], Zhang Jian-ze[1], and Li Xiao-dong[2]

[1] Department of Economy and Management, North China Electric Power University, Baoding, Hebei Province, 071003, China
ahgaochong@yahoo.com.cn
[2] Shunyi Power Supply Company, Beijing Electric Power Corporation, Beijing 101300, China
85103412@qq.com

Abstract. This paper uses the DEA model to analyses the technical efficiency of representative listed companies in China new energy industry representative.Through the analysis and solution for the data on three consecutive years, the results in a certain extent, reflect the development and business efficiency of new energy industry, and provide relevant decision-making information on technical efficiency for new energy industry operation, with strong reliability and practicality.

Keywords: data envelopment analysis（DEA）; new energy; listed company; technical efficiency.

1 Introduction

The contradiction between economic development and resources environment increase sharply, speeding up the adjustment of economic structure and transforming the pattern of resource utilization is imminent. So, accelerating the development and utilization of the new energy such as solar energy, wind energy and so on, achieving energy saving, play a vital role on solving energy and environmental issues and promoting the sustainable development of society. In this context, the new energy industries develop and expand rapidly. It figures out that, China's new energy consumption proportion of total energy consumption by 8.5% in 2007,In the "Long-term development plan for renewable energy" explicitly to 2010 the proportion of renewable energy in China will reach 10%, and 15% by 2020.However, it results the increasing competition. Continuously improving the competitiveness of new energy industries, ensuring the new energy sector has a rapid, sustained and healthy development has important practical significance and research value.

Currently, research on new energy industries is mainly reflected in the industrial system's impact on the industry, new energy technology development issues, and the industrial development strategy and so on, such as Liu Bo (2009), "Development countermeasures of new energy technologies in China ".

* Corresponding author.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 161–168, 2010.

This paper establishes the DEA model to calculate the efficiency indexes of representative listed companies in new energy industry from 2006 to 2008, obtain the depth analysis result, and provide relevant decision-making information on technical efficiency for new energy industry operation, with strong reliability and practicality.

2 Data Envelopment Analysis Model

Data Envelopment Analysis(DEA）, is a non-parametric efficiency frontier analysis method which is widely used to determine technical efficiency. It proposed by well-known operational research experts A. Charnes. W. W. Cooper and E. Rhodes based on the concept of relative efficiency in 1978. The model is used to assess the relative effectiveness (known as DEA efficient) between the same type decision making unit with multiple input and output (abbreviated to DMU). Through the input and output data analysis, based on "Pareto optimal" concept, we determine whether the DMU is DEA efficient, that is , whether the DMU is at the "front surface" of production possibility set essentially. Using the DEA to make efficiency evaluation of DMU, you can get a lot of management information which contains profound economic implications and background in economics.

Given there are n decision making unit, each unit has decision-making has m types of input and s types of output .The input matrix of the $jth\ DMU_{j_0}$ is x_j, the output matrix is y_j.

(1).The CCR model is the first and the most wildly used model of DEA to evaluate the technical efficiency of DMU. In order to judge the effectiveness of decision-making units conveniently, we chose the CCR model with a dimensionless of non-Archimedes:

$$\min[\theta - \varepsilon(e_1^T s^- + e_2^T s^+)] \tag{1}$$

$$s.t.\begin{cases} \sum_{j=1}^{n} x_j \lambda_j + s^- = \theta x_0 \\ \sum_{j=1}^{n} y_j \lambda_j - s^+ = y_0 \\ \lambda_j \geq 0, j = 1,2,...,n \\ s^+, s^- \geq 0. \end{cases} \tag{2}$$

In it, $e_1^T = (1,1,...1)^T \in E^m, e_2^T = (1,1,...1)^T \in E^s$.

ε is dimensionless of non-Archimedes, s^+ and s^- are slack variable.

Assuming that the solution for the linear programming over the optimal is $\sum_{j=1}^{n}\lambda_j^*$, s^{+*} , s^{-*} , θ^* ,and according to the results to determine the efficiency of input-output :

1). If θ^* =1, s^{+*} 、 s^{-*} exist nonzero values, DMU_0 is weak efficient for DEA. If a s^{-*} >0, it says that a certain input index has not been fully used; If a s^{+*} >0,it says that a certain output index has insufficient compared with the maximum output.

2). If θ^* =1, $s^{+*}=s^{-*}$ =0, DMU_0 is efficient for DEA.It says the production activities of DMU_0 is both technology efficiency and scale efficiency, resources are fully utilized,it achieves the largest output effect.

3). If $\frac{1}{\theta^*}\cdot\sum_{j=1}^{n}\lambda_j^*$ =1, DMU_0 is constant returns to scale, the production scale is the best; if $\frac{1}{\theta^*}\cdot\sum_{j=1}^{n}\lambda_j^*$ <1, DMU_0 is increasing returns to scale; if $\frac{1}{\theta^*}\cdot\sum_{j=1}^{n}\lambda_j^*$ >1, , DMU_0 is decreasing returns to scale.

(2).we chooses the BCC model to evaluate Pure Technical Efficiency:

$$\min[\sigma-\varepsilon(e_1^T s^- + e_2^T s^+)] \tag{3}$$

$$s.t.\begin{cases}\sum_{j=1}^{n} x_j\lambda_j + s^- = \sigma x_0\\ \sum_{j=1}^{n} y_j\lambda_j - s^+ = y_0\\ \sum_{j=1}^{n}\lambda_j = 1\\ \lambda_j \geq 0, j=1,2,...,n\\ s^+, s^- \geq 0\end{cases} \tag{4}$$

Economic implications:1)if σ^* =1, s^{+*} 、 s^{-*} exist nonzero values, DMU_0 is weak efficient for DEA.(Pure technology).

2）If σ^* =1, $s^{+*}=s^{-*}$ =0, DMU_0 is efficient for DEA(Pure technology).

(3). DMU_0 scale efficiency equals to the ratio of the technical efficiency and pure technical efficiency.

3 Empirical Study

3.1 Samples and Decision-Making Indexes Selection

There are 56 listed companies in China new energy sector, involved in photovoltaic power generation, biomass power, and wind power equipment manufacturing, solar thermal utilization and so on.

Excluding ST and * ST Series and the companies which losses have occurred from 2006 to 2008 and the disclosure of information is incomplete, then we select the 32 representative companies in the industry as the research samples.(accounted for 57%of new energy listed companies).

As the enterprises are in different professions, in order to research objectively, in this paper, we define company as an organization which uses its available resources, through the production and operation, to make economic profit. The input indicators are: total assets, current assets, total number of employees; the output indicators are: net profit, return on assets. Total assets reflect the overall size of the company level, current assets is used to measure a company's short-term liquidity and short-term solvency of the size. Number of employees reflects the company's technical efficiency and management level, net profit is the core index of the enterprise efficiency, return on assets reflects the ability of enterprises to benefit and the level of their investment income.

3.2 Data Treatment Results

Input the original data into DEAP2.1 software; use the Input-oriented BCC-type model. We can get the companies' technical efficiency, pure technical efficiency, scale efficiency and returns to scale, which is shown in table 1.

3.3 Analysis of Treatment Results

3.3.1 DEA Efficiency Analysis

Observation of Table 1, from 2006-2008, the average of China's new energy listed companies' technology efficiency , pure technical efficiency, scale efficiency are all less than 1, but the value of all three indexes showes increasing trend year by year, which reflect out that China's new energy business stay in a healthy, steady growth, and there is great potential for development. Although the overall efficiency is low currently, it improves steadily. In 2006 the overall technical efficiency (both technical efficiency and pure technical efficiency are valid) are 3, 2007 reaches 5, and 4 in 2008,in which overall technical efficiency for three consecutive years are 2 (Goldwind Technology, Lingguang Industrial).

From Table 2, in 2006 there are only four listed companies reach the pure technical efficiency, 2007 reaches 6, and the same in 2008. In general, the proportion of the total samples is low, which indicate that the pure technical efficiency of the representative enterprises in the new energy industry is not high, but the average increase each year, it also reflects the good momentum of development.

Table 1. New energy listed companies DEA data treatment results from 2006 to 2008

	Listed companies	2006				2007				2008			
		crste	vrste	scale	returns to scale	crste	vrste	scale	returns to scale	crste	vrste	scale	returns to scale
1	Kaidi Power	0.256	0.388	0.660	irs	1.000	1.000	1.000	-	0.801	0.811	0.987	irs
2	Universal Demong	0.071	0.494	0.144	irs	0.285	0.576	0.495	irs	0.318	0.649	0.490	irs
3	Jinshan Shares	0.184	0.431	0.428	irs	0.280	0.353	0.793	irs	0.048	0.185	0.261	irs
4	Goldwind Technology	1.000	1.000	1.000	-	1.000	1.000	1.000	-	1.000	1.000	1.000	-
5	Silver Star Energy	0.023	0.365	0.063	irs	0.107	0.398	0.269	irs	0.248	0.440	0.564	irs
6	Antai Technology	0.158	0.219	0.719	irs	0.294	0.298	0.987	irs	0.329	0.344	0.957	drs
7	Great Wall Electrical	0.027	0.132	0.207	irs	0.153	0.191	0.801	irs	0.085	0.207	0.413	irs
8	Tibet Mining	0.125	0.417	0.299	irs	0.312	0.423	0.738	irs	0.376	0.762	0.494	irs
9	Leshan Power	0.308	0.684	0.451	irs	0.514	0.580	0.887	irs	0.295	0.300	0.981	irs
10	Linggang Industrial	1.000	1.000	1.000	-	1.000	1.000	1.000	-	1.000	1.000	1.000	-
11	Taihao Technology	0.098	0.190	0.518	irs	0.209	0.227	0.919	irs	0.189	0.195	0.969	irs
12	Jiuding New Material	0.159	0.751	0.212	irs	0.232	0.597	0.389	irs	0.177	0.795	0.223	irs
13	Guanghui Shares	0.031	0.234	0.133	irs	0.809	0.922	0.877	drs	0.924	1.000	0.924	drs
14	Zhongcai Technology	0.124	0.269	0.462	irs	0.357	0.359	0.994	drs	0.390	0.403	0.967	drs
15	Huayi Electric	0.031	0.577	0.055	irs	0.592	0.630	0.938	irs	0.471	0.541	0.870	irs
16	Wolong Electric	0.134	0.186	0.721	irs	0.435	0.532	0.818	drs	0.453	0.544	0.834	drs
17	CNAC Sanein	0.104	0.556	0.187	irs	0.262	0.419	0.625	irs	0.177	0.390	0.454	irs
18	Furi Shares	0.081	0.088	0.928	irs	0.131	0.176	0.745	drs	0.110	0.116	0.947	drs
19	Huitong Energy	0.193	1.000	0.193	irs	1.000	1.000	1.000	-	1.000	1.000	1.000	-
20	Xiangdian Shares	0.063	0.101	0.622	irs	0.128	0.130	0.990	irs	0.077	0.118	0.652	irs
21	Aerospace Electrical	0.189	0.219	0.866	irs	0.190	0.234	0.814	irs	0.115	0.268	0.429	irs
22	Haina shares	0.258	0.367	0.704	irs	0.805	1.000	0.805	drs	0.010	0.088	0.110	irs
23	Shenhua Holdings	0.407	0.945	0.431	irs	0.740	0.932	0.794	irs	0.493	0.568	0.868	irs
24	Wanneng Power	0.117	0.172	0.680	irs	0.116	0.136	0.854	irs	0.037	0.128	0.289	irs
25	Meiling Power	0.011	0.174	0.065	irs	0.017	0.164	0.105	irs	0.046	0.180	0.253	irs
26	Jiangsu Sunshine	0.074	0.088	0.839	irs	0.213	0.258	0.827	drs	0.160	0.205	0.784	drs
27	New Huaguang	0.029	0.564	0.052	irs	0.044	0.554	0.079	irs	0.059	0.487	0.120	irs
28	Shenneng Shares	1.000	1.000	1.000	-	1.000	1.000	1.000	-	0.869	0.982	0.885	drs
29	Baoxing Energy	0.607	0.678	0.895	irs	0.813	0.887	0.916	irs	1.000	1.000	1.000	-
30	Chitianhua	0.365	0.397	0.921	irs	0.564	0.595	0.949	drs	0.357	0.359	0.994	irs
31	Tianwei Electric	0.213	0.250	0.851	irs	0.511	0.535	0.956	drs	0.855	1.000	0.855	drs
32	Fanta Group	0.060	0.404	0.149	irs	0.126	0.378	0.335	irs	0.088	0.410	0.214	irs
	Mean	0.228	0.446	0.511	—	0.435	0.642	0.676	—	0.488	0.693	0.703	—

Note：（1）crste- technical efficiency; vrste-pure technical efficiency; scale- scale efficiency;
（2）irs- increasing returns to scale; drs- decreasing returns to scale;-- constant returns to scale;
（3）The original data is extracted from Sina Finance companies summary annual report and financial indexes from 2006 to 2008.

Table 2. DEA efficiency analysis

		Overall technical efficiency	Pure technical efficiency
2006	Number	3	4
	Percentage	9.38%	12.50%
2007	Number	5	6
	Percentage	15.63%	18.75%
2008	Number	4	6
	Percentage	12.50%	18.75%

3.3.2 Returns to Scale Analysis

From Table 1, in the studied samples, but only a few reached the point of constant returns to scale, there are just 4 in 2008, accounting for the low proportion of study sample. In 2006, the listed companies except the constant returns to scale, all the other companies are at the stage of increasing returns to scale, it illustrates that under the support of the national industrial policy, New energy industry stay in a rapidly developing stage. Seeing from table 3, more than 60.61% of new energy listed companies are at the stage of increasing returns to scale, it points out that the new energy companies can increase production factor and expand the production scale to obtain a greater output

Table 3. Returns to scale analysis

		Increasing returns to scale	Constant returns to scale	Decreasing returns to scale	Sum
2006	Number	29	3	0	32
	Percentage	90.63%	9.37%	0	100%
2007	Number	19	5	8	32
	Percentage	59.38%	15.62%	25.00%	100%
2008	Number	20	4	8	32
	Percentage	62.50%	12.50%	25.00%	100%

The selected sample enterprises should take appropriate measures to adjust the size of enterprise under the circumstances.If a company stay in the increasing returns to scale in the business, assets investment should be increased to enhance output levels; if a company stay in the decreasing returns to scale, it should reduce the business assets scale appropriately, adjust the capital structure, strengthen management and improve the efficiency of input and output, so as to adjust to the optimum size.

3.3.3 Improvement of the Theoretical Value Analysis

Table 4 lists the slack variable measured values of the increasing returns to scale enterprises , we can improve the enterprises which technical efficiency is less than 1, so that it can achieve the target improvements in the theoretical value. Seeing from the table 4, most of the increasing returns to scale enterprises have redundancy in current assets and number of staff, while the return on assets output is shortfall. Take Huayi Electric as example, the company's current assets was 926,576,600 yuan, 905 employees, return on assets was 18.26%, its technical efficiency was 0.471 in 2008, redundancy for the input current assets is 228,609,130 yuan, for the staff is151,and the current assets improvement value is 207,808,400 (θx_2 -228,609,130) yuan; number of staff improvement value is 276(θx_3 -151) ; return on assets was 1.49% less than the maximal output, the return on assets improvement value is 19.75%(y_2 +1.49%) .In this way, we can get the theoretical improvement value, to provide relevant information for company decision-making. In addition, enterprises

Table 4. Slack variable measured value of the increasing returns to scale enterprises

Listed companies	s_1^- s_2^- s_3^-	s_1^+ s_2^+
Kaidi Power	403698.684 41502.984 0	0 11.771
Universal Denong	0 3869.191 369	630.780 3.350
Jinshan Shares	50166.717 0 0	5464.170 18.024
Silver Star Energy	0 0 58	8214.543 10.328
Great Wall Electrical	0 8590.883 913	252.800 9.470
Tibet Mining	10915.680 0 804	5946.490 16.750
Leshan Power	1805.047 0 0	0 7.292
Taihao Technology	0 13248.672 0	0 11.229
Jiuding New Material	0 0 1153	5433.574 12.529
Huayi Electric	0 22866.913 151	0 1.490
CNAC Sanxin	0 5179.299 903	0 4.860
Xiangdian Shares	0 13212.562 365	0 11.803
Aerospace Electrical	4224.250 0 0	4546.145 16.749
Haima shares	0 690.982 347	4003.530 10.580
Shenhua Holdings	163554.757 61335.655 0	404.820 10.710
Wanneng Power	51440.907 0 0	6786.787 20.995
Meiling Power	0 0 195	5206.481 15.942
New Huaguang	0 5239.969 332	3390.910 10.420
Chi Tianhua	0 0 0	0 8.431
Fanta Group	0 0 316	3342.522 10.105

can also take measures such as introduce advanced technology, improve management and efficiency to make the output level to increase to enable enterprises to achieve the desired scale invariant point.

4 Conclusions

This paper uses DEA model to analyse the technical efficiency of representative listed companies in China new energy industry, So as to get the current development operating conditions of new energy industry. The results show that,the overall technical efficiency of the representative listed companies in new energy industry for three consecutive years on average is less than 1, overall efficiency is not high. But the trend is rising year by year, and most of these 32 companies are in the stage of increasing returns to scale.So that in the current background of new energy development and low carbon policies , companies should extensively increase business investment, adjust industrial structure appropriately, in order to obtain the optimal output.

DEA models is used extensively in production management effectiveness analysis, and this method has high validity and reliability. This paper solve and calculate the model, the results can to some extent reflect the current situation of new energy market business operations, provide useful information and guide observations for decision-makers.

Acknowledgment

Project Supported by the Young Teacher Foundation of North China Electric Power University(200911027).

References

1. Wei, Q.L.: Data Envelopment Analysis (DEA). Science Bulletin. 17 (2000)
2. Wei, L., Yue-feng, N.: A DEA-based Analysis into the input and output of the major listed private enterprises of Forbes top 100 Chinese ones in 2008. Technological Economics & Management Research 3, 14–16 (2009)
3. Xian-zhong, M., Bing-yi, L.: The Research of New Energy and Renewable Energy Development and Industrialization (2009)
4. Bo, L.: Development and Countermeasures of new energy technologies in China. Journal of Liaoning University of Technology (Social Science Edition) (2), 30–33 (2009)
5. Niu, D.-x., Wang, X.-m., Gao, C.: The combination evaluation method based on DEA and SVM. In: Proceeding of 2008 International Conference on Machine Learning and Cybernetics, vol. (2), pp. 845–850 (2008)

Research on Differential Evolution Algorithm Based on Interaction Balance Method for Large-Scale Industrial Processes of Fuzzy Model

He Dakuo, Zhao Yuanyuan, Wang Lifeng, and Chang Hongrui

Key Laboratory of Process Industry Automation, Ministry of Education,
Northeastern University, 110819, Shenyang, China
hedakuo@mail.neu.edu.cn

Abstract. As a result of slow perturbations, the mathematical model of an actual system is difficult to be accurate. So when optimizing large-scale industrial process, the mathematical model and the actual system does not match, that is model-actual difference. Large-scale industrial process optimization based on fuzzy model is an effective way of this issue. However, the optimization model is the process of establishing a non-linear programming model. So, differential evolution algorithm is studied this paper to solve the problems of large-scale industrial processes optimization based on fuzzy models. To solve model-actual differences is mainly to solve fuzzy nonlinear problems: firstly, differential evolution algorithm is studied for solving fuzzy nonlinear problems in this paper. Then the combination of fuzzy nonlinear problems and the interaction balance method coordinated approach of large-scale industrial processes is proposed. Last, simulation results show the validity of the method which this paper studies.

Keywords: fuzzy nonlinear programming; fuzzy model; large-scale industrial process; interaction balance method; differential evolution algorithm.

1 Introduction

In order to solve the model-reality differences, the Reference[1] proposed using the interaction balance coordination method of large-scale industrial processes with fuzzy parameters. It converts the model of large-scale industrial optimization processes into fuzzy nonlinear programming problems and solves them. There are three typical methods to solve the fuzzy nonlinear programming problems of the current references. Firstly, Pan Yu[2]proposed a method to solve the first type of fuzzy nonlinear programming problems, that is, to solve such problems using Zoutendijk method. Tang Jiafu and Wang Dingwei[3]have discussed the fuzzy optimum method based on genetic algorithm to solve the quadratic programming problems and the nonlinear programming problems with fuzzy object and fuzzy resource constraints. Gu Jiachen[4] provided a way of converting the fuzzy nonlinear programming problems into accurate ones to solve fuzzy nonlinear

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 169–176, 2010.

programming problems. The ways of solving the fuzzy nonlinear programming problems above were always about converting the fuzzy nonlinear programming problems into accurate ones, then using the common methods to solve the fuzzy problems. However, such ways have large limitations to some extent, because the large-scale industrial processes optimization model itself is an approximate one, if the fuzzy model is converted into accurate one, there will be serious errors and cause the model deviating reality more serious. So this paper proposes a way of using differential evolution algorithm to solve fuzzy nonlinear programming problems.

2 The Open-Loop Ibm Based on Fuzzy Model for Large-Scale Industrial Processes

2.1 Description of the Algorithm about the Open-Loop IBM

Interaction balance method is a basic method for solving the steady-state optimization problem of large-scale industrial processes. The basic idea is to cut off the interactions of every subsystem, and the interaction inputs are referred as independent optimization variables of the local decision units. Considering the model-reality differences, when the model coefficients of subprocesses are treated as fuzzy numbers, the IBM can be derived[1-5-6].

The local decision units' problems are:

$$(FLP_i)\begin{cases} for\ the\ given\ \lambda, \\ to\ find\ \hat{c}_i(\lambda)\ and\ \hat{u}_i(\lambda), \hat{y}_i(\lambda), such\ that \\ (\hat{c}_i(\lambda), \hat{u}_i(\lambda), \hat{y}_i(\lambda)) = \arg\min L_i(c_i, u_i, y_i, \lambda) \\ s.t.\ \ y_{ik} - \tilde{F}_{ik}(c_i, u_i) = \tilde{0}\ (k = 1,2,\cdots,n) \\ \qquad g_{ij}(c_i, u_i, y_i) \lesssim 0 \\ \qquad (c_i, u_i, y_i) \in CUY_i \end{cases} \tag{1}$$

The coordination problem is:

Find $\hat{\lambda}$, such that

$$\hat{u}(\hat{\lambda}) = H\hat{y}(\hat{\lambda}) \tag{2}$$

Where the variables u_i, y_{ik} and c_i represent the *i* th subprocesses inputs, outputs and control variables respectively and they are accurate ones.

The formula $y_{ik} - \tilde{F}_{ik}(c_i, u_i, \tilde{a}_i) = \tilde{0}\ (k = 1, \cdots, n_i)$ is the fuzzy equality constraints of ith subprocesses which the coefficients are fuzzified. The expression $g_{ij}(c_i, u_i, y_i) \lesssim 0$ is the inequality constrains of the *i* th subprocesses. λ is the Language multiplier vector of the interaction constraints *u=Hy* with the interaction matrix *H*,

and the interaction equations are put into the objective function as the revision. The Language function is:

$$L(c,u,y,\lambda) = Q(c,u,y) + \lambda^T (u - Hy) \tag{3}$$

The equation (3) can be divided into the following parts:

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \left\{ Q_i(c_i,u_i,y_i) + \lambda_i^T u_i - \sum_{j=1}^{N} \lambda_j^T H_{ji} y_i \right\} \tag{4}$$

So the update equation of λ is:

$$\lambda^{k+1} = \lambda^k + \varepsilon_k \nabla D(\lambda^k) \tag{5}$$

Where ε_k is step length coefficient and $\nabla D(\lambda) = \hat{u} - H\hat{y}$. Then the coordination problem can be solved effectively by the gradient search optimization method.

Under the steady state, the mathematical model of every subprocesses of the large-scale industrial process is linear; it can be described by a group of linear algebra equations.

$$y_{ik} = F_{ik}(c_i,u_i) = a_{ik1}c_{i1} + \cdots a_{ikp}c_{ip} + a_{ikp+1}u_{i1} + \cdots a_{ikp}u_{ip} \tag{6}$$

Where $a_{ik} = (a_{ik},\cdots,a_{ikp+n_i})$ is the coefficient vector of the k th equation of the i the subprocesses model. Considering the differences between model and reality, and these differences are usually those the coefficients of the models which are inaccurate, so the coefficients $a_{ik} = (a_{ik},\cdots,a_{ikp+n_i})$ can be treated as fuzzy number $\tilde{a}_{ik} = (\tilde{a}_{ik},\cdots,\tilde{a}_{ikp+n_i})$. Then the mathematical models then the mathematical models converted into the following fuzzy equation:

$$y_{ik} = \tilde{F}_{ik}(c_i,u_i,\tilde{a}_i) = \tilde{a}_{ik1}c_{i1} + \cdots \tilde{a}_{ikp}c_{ip} + \tilde{a}_{ikp+1}u_{i1} + \cdots \tilde{a}_{ikp}u_{ip} \tag{7}$$

Where $u_i = (u_{i1},\cdots,u_{n_i})$, $y_i = (y_{i1},\cdots,y_{n_i})$ and c_i and i th subprocesses inputs, outputs and control c_i and i th subprocesses inputs, outputs and control variables, they' re accurate variables. $\tilde{a}_{ik} = (\tilde{a}_{ik},\cdots,\tilde{a}_{ikp+n_i})$ are *L-R* type fuzzy numbers, they can be expressed as follows:

$$\tilde{a}_{ik1} = (a_{ik1};\underline{a_{ik1}},\overline{a_{ik1}}),\cdots,\tilde{a}_{ikp+n} = (a_{ikp+n};\underline{a_{ikp+n}},\overline{a_{ikp+n}}) \tag{8}$$

So the equations with fuzzy numbers can be dealt with by the method above, then the fuzzy equality constraints are converted into the following ones:

Fuzzy inequality constraints is $g_{ij}(c_i,u_i,y_i) \tilde{\leq} \tilde{0}$, *i=1,2, ⋯ ,n, j=1,2, ⋯ ,n, m* can be treated as such ones with elastic restrictions. And inequality constraint has the elastic property $l=(l_1,l_2,\cdots,l_m)$.

2.2 Using Differential Evolution Algorithm to Solve Fuzzy Nonlinear Programming Problems in Local Decision Units

Differential evolution algorithm is proposed to solve the first type fuzzy nonlinear programming problem in this paper, and the concrete form of the first type fuzzy nonlinear programming is as follows:

$$\begin{cases} \min(\max) f(x) \\ s.t.\ \ g_i(x) \tilde{\leq} b_i, i = 1, \cdots, m \\ \quad h_j(x) \cong d_j, j = 1, \cdots, n \end{cases} \tag{9}$$

And the elastic limits of constraints can be expressed by $l=(l_1,l_2,...,l_m,l_{m+1},...,l_n)$.

The objective function and constraints functions are fuzzified as:
The membership function of inequality constraint is expressed as follows:

$$\mu_{g_i}(x) = \begin{cases} 1 & g_i \leq b_i \\ 1 - (\dfrac{g_i - b_i}{l_i})^r & b_i < g_i < b_i + l_i \\ 0 & g_i > b_i + l_i \end{cases} \tag{10}$$

The membership function of equality constraint is described as:

$$\mu_{h_j}(x) = \begin{cases} 0 & h_j \leq d_j - l_j^+ \\ 1 - (\dfrac{d_j - h_j}{l_j^+})^r & d_j - l_j^+ < h_j \leq d_j \\ 1 - (\dfrac{h_j - d_j}{l_j^+})^r & d_j < h_j \leq d_j + l_j^+ \\ 0 & d_j > d_j + l_j^+ \end{cases} \tag{11}$$

The membership function of objective function is:

$$\mu_0(x) = \begin{cases} 0 & f(x) \leq z_0 - p_0 \\ (\dfrac{z_0 - f(x)}{p_0})^r & z_0 - p_0 < f(x) \leq z_0 \\ 1 & f(x) > z_0 \end{cases} \tag{12}$$

The elastic limit of every soft constraint is: $l=(l_1,l_2,\cdots,l_m, l_{m+1}, \cdots,l_{n+1})$.

An improved differential evolution algorithm based on the elementary differential evolution is proposed in this paper, that is, Local Search Differential Evolution. The main difference between DE and LSDE is that :the LSDE algorithm starts from a initial solution X, then searches a solution $X`$ better than X in the epsilon neighbor of

X. If the solution $X`$ is better than $X$, then $X$ is replaced by $X`$, after that the local search should be going on, else the search is over.

The specific procedure that using LSDE to solve fuzzy nonlinear programming problems is as follows:

Step 1(initialization): The population size N, initial mutation factor F_0, hybrid probability *Max Gens* are determined. And an uniform distribution of initial population is generated randomly. N is between 20 and 50. A local search is done every 20 generations, and the local search times is m. Then the maximum iterations is SS_{max} in each local search.

Step 2 : The membership functions of each individual in the initial population is calculated ,then evaluate the individuals .The fitness function is

$$f(x_i) = \frac{f(x_i)}{\sum_{i=1}^{N} f(x_i)} = \mu_{\min}(x_i) = \min\{\mu_{\min}(x), i = 0,1,2,...,n\}$$

Find the individual having the best fitness value and write it as X_{best} and its fitness value as F_{best}. And $index(X_{best})$ represents the individual index in the population. Then set up the evolutionary generations G=1.

Step 3(mutation): Execute mutation according to the formula:

$$V_{i,G+1} = X_{r_1,G} + F_G \times (X_{r_2,G} - X_{r_3,G})$$

where $F_G = e^{-G/A \times MaxGens} \times B \times F_{G-1} \times (1 - F_{G-1})$, A=3, B=4, then get the vector $V_{i,G+1}$.

Step 4(crossover): The crossover is executed according to the formula as follow:

$$u_{ji,G+1} = \begin{cases} V_{ji,G+1}, & if\ (randb(j) \le CR)\ or\ (j = rnbr(i)) \\ X_{ji,G+1}, & if\ (randb(j) > CR)\ and\ (j \ne rnbr(i)) \end{cases}$$

($j=1,2,\cdots,D$). $U_{i,G+1}$ is a new vector obtained in the crossover operation.

Step 5(selection): Execute selection and calculate the fitness value of each individual after crossover operation , then choose the minimum one as the individual of next generation $X_{i,G+1}$.

Step 6(comparison)**:** Compare $f(x_{i,G+1})$ with F_{best}, if $f(x_{i,G+1}) < F_{best}$, then let $X_{best} = x_{i,G+1}$, $F_{best} = f(x_{i,G+1})$ and $index(X_{best}) = i$.

Step 7: If $G = 20 \times m$, then set $m = m+1$, and go to the next step ,else go to step 10.

Step 8(Local search): Generate a small population around the X_{best} through mutation, crossover and selection, then execute local search k times and find the $x_b^{(k)}$ and $f(X_b^{(k)})$.

Step 9: Compare the fitness value $f(X_b^{(k)})$ with F_{best}. If the fitness value $f(X_b^{(k)})$ of $x_b^{(k)}$ is better than F_{best}, then let $X_{best} = x_b^{(k)}$, $F_{best} = f(X_b^{(k)})$ and generate a random integer $j \in [1, N]$, $index(X_{best}) = j$, $X_{i,G+1}$.

Step 10: G=G+1, if G < *MaxGens*, then turn to step3, else continue.

Step 11: Output X_{best} and F_{best}.

2.3 Simulation Research

Example [7] is composed of three interacted subprocesses .Its structure scheme is as follows:

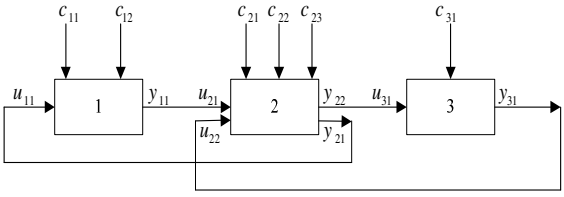


Fig. 1. The system structure

Simulation results are expressed in Table 1.

Table 1. Simulation Result of Open-loop Interaction Balance Method Based on Fuzzy Model

Model	Method	Model objective function	Practical objective function	Iterations	Constraints
Example	Accurate parameters	2.4833	2.4048	44	All constraints preserved
	Fuzzy parameters in the Reference[6]	2.4269	2.0543	44	All constraints preserved
	Fuzzy parameters in this paper	2.0285	2.0325	40	*All constraints preserved*

Some conclusions can be gotten from Table 1, the method proposed in this paper is better than the method in Reference[4]. The objective function value based on interaction balance method for large-scale industrial process of fuzzy model is much smaller than the value based on interaction balance method for large-scale industrial process with accurate numbers, and the iterations are not increasing. This proves it is feasible to deal with the model-reality differences in steady state optimization for large-scale industrial processes, and there are certain high performances by using fuzzy method in the open-loop interaction balance method.

3 The Global Feedback IBM Based on Fuzzy Model for Large-Scale Industrial Processes

3.1 Description of the Algorithm about the Global Feedback IBM

The global feedback IBM was proposed by Poland scholar Findeisen firstly, the aim is to conquer the suboptimality of the results brought by model-reality differences. And the basic idea is to collect the steady-state messages of the interactive variables from the actual process, then send the steady-state messages into the coordination and modify the optimized result. So the result can approach the actual process optimized result[8-9].

The local decision units' problems are :

$$(FLP_i)\begin{cases} for\,the\,given\,\lambda \\ to\,find\,(\hat{c}_i(\lambda),\hat{u}_i(\lambda)) = arm\min L_i(c_i,u_i,y_i,\lambda) \\ s.t.\ \ y_{ik} - F_{ik}(c_i,u_i) = 0 (k = 1,\cdots,n_i) \\ \quad g_{ij}(c_i,u_i,y_i) \tilde{\leq} 0 \\ \quad (c_i,u_i,y_i) \in CUY_i \end{cases} \tag{13}$$

The coordination problem is:

$$(CP)\begin{cases} to\ find\ a\ coordinating\ \ \text{variable}\ \ \hat{\lambda} = (\hat{\lambda}_1,\cdots,\hat{\lambda}_N) \\ such\ that\ \hat{u}(\hat{\lambda}) = u^*(\hat{\lambda}) = HK^*(\hat{c}(\hat{\lambda})) \end{cases} \tag{14}$$

The iterative policy to solve the problem above is :

$$\lambda^{k+1} = \lambda^k - a[\hat{u}(\lambda^k) - u^*(\lambda^k)] \tag{15}$$

Where a is the iterative step length. The initial point can be chosen based on the optimized coordinating variable λ^0 of the model.

3.2 Simulation Research

The example is the same as section2.3. The simulation results are in Table 3.

Table 2. The Simulation Results of Global Feedback Interaction Balance Method Based on Fuzzy Model

Model	Method	Practical objective function	Iterations	Constraints
Example	Accurate parameters	2.0411	112	All constraints preserved
	Fuzzy parameters in the reference[6]	2.0411	30	All constraints preserved
	Fuzzy parameters in this paper	2.0210	30	All constraints preserved

The conclusions gotten from Table 3 is that the method proposed in this paper is better than the method in Reference[4]. The on-line iterations based on global feedback interaction balance method for large-scale industrial process of fuzzy model are much fewer than the method with accurate numbers. So the disturbance to the actual process reduces greatly, and the objective function value decreases to some extent.

4 Conclusion

The differential evolution algorithm is introduced in this paper to solve the fuzzy nonlinear programming problems which is applied to the interaction balance method

of large-scale industrial processes optimization. Then the interaction balance method model is converted into fuzzy nonlinear model and differential evolution algorithm is used to solve it. The simulation results of the typical example about the large-scale industrial processes optimization have proved the method this paper proposed and solved the model-reality differences.

Acknowledgment

This work is supported by the National Nature Science Foundation of China under Grant 61004083 and 973 Project No. 2009CB320601 and the Fundamental Research Funds for the Central Universities N090404005.

References

[1] Gu, J.C., Wan, B.W., Guan, X.H.: Interaction Balance Method for Large-Scale Industrial Processes with Fuzzy Models, vol. 28(4), pp. 569–574 (2002)
[2] Wan, B.W., Huang, Z.L.: On-line Steady State Hierarchical Optimizing Control of Large-scale Industrial Processes. Science Press, Beijing (1998) (in Chinese)
[3] Zhang, Z.K.: Applications of Fuzzy Maths in Automatics. Science Press, Beijing (1997)
[4] Findeisen, W., Bailey, F., Brdys, M., et al.: Control and Coordination in Hierarchical Systems. John Wiley and Sons, London (1980)
[5] Gu, J.C.: Steady State Hierarchical Optimizing Control for Large-Scale Industrial Processes With Fuzzy Parameters. Xi'an Jiaotong University, Xi'an (2001)
[6] Shao, F.Q., Roberts, P.D.: A price correction mechanism with global feedback for hierarchical control of steady-state systems. Large Scale Systems 4(2), 67–80 (1983)
[7] Li, L.Z., Qin, S.Y., Wan, B.W.: A Novel Intelligent Optimizing Method for Large-Scale Steady-State Systems with Fuzzy Parameters. Journal of Systems Engineering 7(2), 1–11 (1992)
[8] Pan, Y.: An algorithm of the first class of fuzzy nonlinear progrmming. Journal of Nanjing University of Technology 12(3), 30–36 (1990)
[9] Tang, J.F., Ding, D.W.: An interactive approach based on GA for a type of quadratic programming problems with fuzzy objective and resoures. Computer & Operations 24(5), 43–422 (1997)

Fuzzy Control System Design for Solar Energy with PCM Tank to Fresh Air Conditioning

Jun Yang, Ailin Xiao, and Lili Wang

Beijing University of Posts and Telecommunications
Beijing, China
Jun Yang yjun@bupt.edu.cn
{Ailin Xiao,Lili Wang}xooostrugle@yahoo.cn

Abstract. With the help of Solar-collector and PCM (Phase Change Material) tank, an experimental system has been set up. In this system, the framework of the system is based on fuzzy algorithm. To provide a better precision of the fresh-air temperature, the system works in three different modes which need three different control rules. In this paper, these three modes are introduced firstly, and then a design of the three fuzzy controllers is described in detail. Meanwhile, the automatic transition process of the three modes in different conditions is studied. Finally, through the analysis of experimental data, it is proved that the system can provide stable fresh air temperature according to different setting temperature.

Keywords: fresh air; solar energy; fuzz Control; PCM.

1 Introduction

The fresh-air system is a common phenomenon in developed countries. In some countries, it is especially regarded as one of the necessary equipment installed.

Compared with traditional air conditioner, it will save much more energy. Moreover, solar energy is an ideal green energy as renewable energy. Nowadays, researching on the solar thermal applications has become a hot topic and a common trend.

Another big advantage using the solar energy to heat fresh air into the room is to improve room-air quality without a lot of heat loss which produces much influence on the existing air-conditioning system.

An experimental system has been built for research. It uses solar as the only source, combines PCM tank to be efficient energy storage, uses CAN-bus models [1] to realize the field communication between each nodes and PC, and uses LabWindows/CVI to build user interface and ActiveX component technology to call the fuzzy control function of MATLAB for online-debugging [2].

In this paper, the research platform is built in Beijing which has abundant solar energy and the longer year-round sunshine. According to the practical application, the fresh air system should work in three different modes which are Mode 1, Mode 2 and Mode 3. The transition process of the modes to each other can be controlled by Electric Valves.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 177–184, 2010.

Mode 2 is the solar energy for both fresh air and PCM tank. The operation can be seen from Fig. 1.

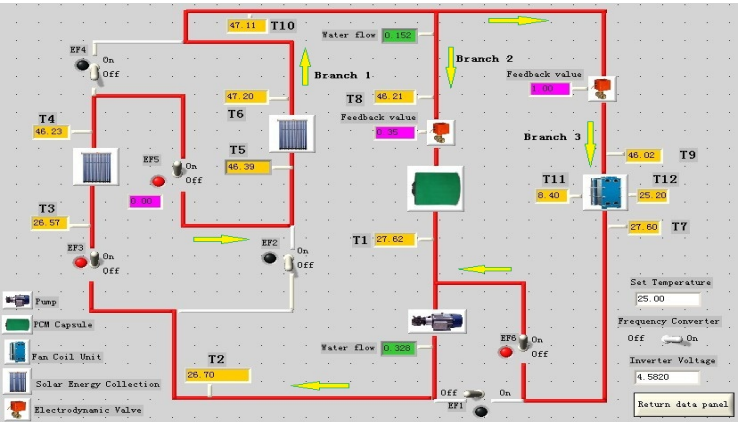


Fig. 1. The operation of Mode 2

Mode 1 is the solar energy only for fresh air. Comparing with Mode 2, it is made up of branch 1 and branch 3, but without branch 2. Take Fig. 1 as a reference, the current is along the time direction.

Mode 3 is PCM tank for fresh air. Comparing with Mode 2, it is made up of branch 2 and branch 3, but without branch 1. In Fig. 2 as a reference, the current is against the direction of time. The key point is that, even if sunset or in the night, the stable operation of the system can be still achieved just by automatic switching to the Mode 3. It contributes to prolonging the heating time. The whole process relies on the system's own judgments, automatic selection and self-regulation.

2 PCM Selection

Koca,Ahmet stored the heat energy of solar collectors in the Thermal Storage Using CaC12·6H2O as phase change material, and the energy efficiency and radiation efficiency is respectively 45% and 2.2%. It is proved that advantage of energy saving and environment protection for the phase change material [3~9].

Additionally, in order to improve the PCM capability, the honeycomb structure is used by the PCM tank. This structure can eliminate the pressure which is caused by phase-change materials volume expanding [10]. Simple structure is shown in Fig. 2.

The PCM is used for the solar energy storage in the experimental system.

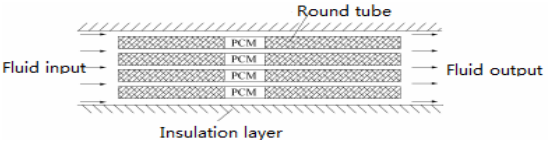


Fig. 2. The PCM thermal storage unit by round tube packaging

3 Fuzzy-Controller Design

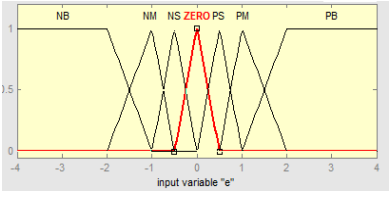
3.1 The Variables of Fuzzy Controllers: Double Inputs and Single Output

Taking error e and error change rate ec as the inputs, and a voltage which used to adjust the frequency converter as the output, the manual control experiments must be done before the fuzzy controller design. The regulation law of the system can be got through the manual experiments.

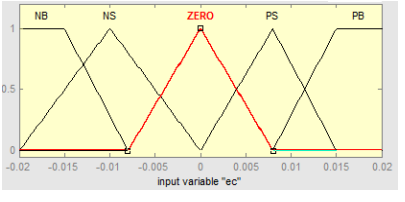
3.2 The Establishment of Membership Functions and Fuzzy Control Rules

Each mode has its own operating conditions and characteristics, which requires its own membership functions and control rules. In order to provide a better precision of the fresh-air temperature and accelerate the response time, the system should work in three different modes.

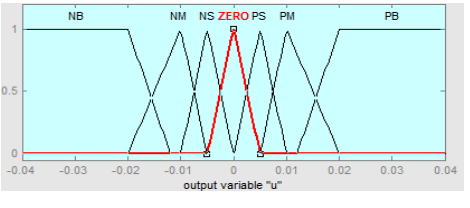
According to the Mode 1 and the data obtained from the manual experiments, we could get the range of e is [-4, 4], and the distribution after fuzzification is [-4, -2, -1, -0.5, 0, 0.5, 1, 2, 4]. The range of ec is [-0.02, 0.02], and the distribution after fuzzification is [-0.02, -0.015, -0.01, -0.005, 0, 0.005, 0.01, 0.015, 0.02]. The range of △u is [-0.04, 0.04], and the distribution after fuzzification is [-0.04, -0.02, -0.012, -0.01, -0.005, 0, 0.005, 0.01, 0.012, 0.02, 0.04]. It is noteworthy that membership functions are mainly distributed near zero, which contributes to improving the control precision. All the distributions are shown in Fig. 3.



(a) Membership function of error e



(b) Membership function of error change rate ec



(c) Adjustment voltage for frequency converter △u

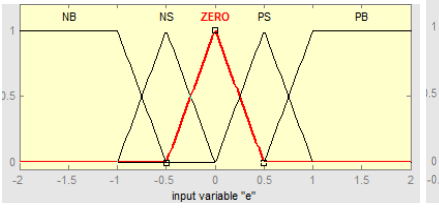
Fig. 3. Membership function of error e, error changer rate ec and adjustment Voltage for frequency converter △u

The fuzzy control rules are shown in Table 1.

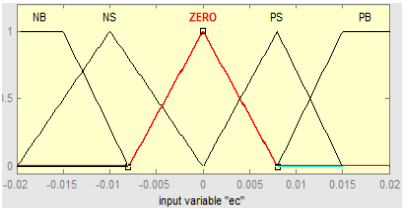
Table 1. The Fuzzy Control Rules Of Mode 1

ec \ e, △u	NB	NM	NS	ZERO	PS	PM	PB
NB	PB	PM	PM	PM	PS	ZERO	NS
NS	PB	PM	PM	PS	ZERO	NS	NS
ZERO	PM	PS	PS	ZERO	NS	NS	NM
PS	PS	PS	ZERO	NS	NM	NM	NB
PB	PS	ZERO	NS	NM	NM	NM	NB

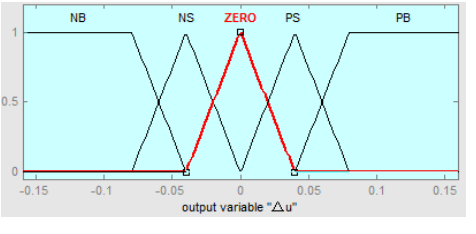
For the Mode 2, we could get the range of e is [-2, 2], and the distribution after fuzzification is [-2, -1, -0.5, 0, 0.5, 1, 2]. The range of ec is [-0.02, 0.02], and the distribution after fuzzification is [-0.02, -0.015, -0.01, -0.005, 0, 0.005, 0.01, 0.015, 0.02]. The range of △u is [-0.15, 0.15], and the distribution after fuzzification is [-0.15, -0.1, -0.05, 0, 0.05, 0.1, 0.15]. All the distributions are shown in Fig. 4.



(a) Membership function of error e



(b) Membership function of error change rate ec



(c) Adjustment voltage for frequency converter △u

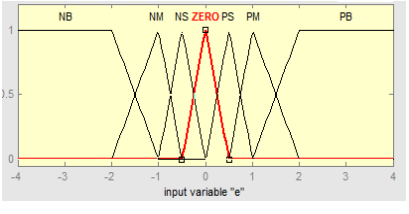
Fig. 4. Membership function of error e, error changer rate ec and adjustment voltage for frequency converter △u

The fuzzy control rules are shown in Table 2.

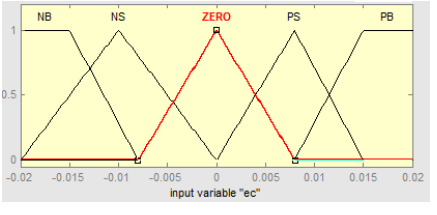
Table 2. The Fuzzy Control Rules Of Mode 2

△u / e / ec	NB	NS	ZERO	PS	PB
NB	PB	PB	PS	PS	ZERO
NS	PB	PB	PS	ZERO	NS
ZERO	PB	PS	ZERO	ZERO	NS
PS	PS	ZERO	ZERO	NS	NB
PB	ZERO	ZERO	NS	NB	NB

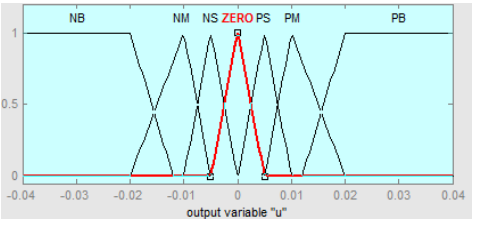
For the Mode 3, we could get the range of e is [-4, 4], and the distribution after fuzzification is [-4, -2, -1, -0.5, 0, 0.5, 1, 2, 4]. The range of ec is [-0.02, 0.02], and the distribution after fuzzification is [-0.02, -0.015, -0.01, -0.005, 0, 0.005, 0.01, 0.015, 0.02]. The range of △u is [-0.2, 0.2], and the distribution after fuzzification is [-0.2, -0.08, -0.06, -0.04, 0, 0.04, 0.06, 0.08 0.04]. For the system characteristics of inertia and delay, the membership function of △u should distribute near the positive and negative 0.05. It can help to improve the stability. All the distributions are shown in Fig. 5.



(a) Membership function of error e



(b) Membership function of error change rate ec



(c) Adjustment voltage for frequency converter △u

Fig. 5. Membership function of error e, error changer rate ec and adjustment voltage for frequency converter △u

The fuzzy control rules are shown in Table 3.

Table 3. The Fuzzy Control Rules Of Mode 3

ec \ e	NB	NS	ZERO	PS	PB
NB	PB	PM	PS	PS	NS
NS	PB	PM	PS	ZERO	NM
ZERO	PM	PS	ZERO	ZERO	NM
PS	PM	ZERO	ZERO	NS	NB
PB	PS	ZERO	NS	NM	NB

4 Experimental Data and Analysis

4.1 Different Setpoint of Fresh Air Temperature Control

According to the practical application, the fresh air system should provide different fresh air temperature for users to choose. And the system must adjust automatically with the change of the setting temperature and has fast response speed. Meanwhile, the fuzzy control system must resist external interference and has basic robustness.

For providing fresh air of different temperature, the Mode 1 is taken as an example. The automatically controlled effect of different specified temperatures can be seen from Fig. 6. During that time, the outside air temperature was about 7℃. The temperature of the water into the fan coil was between 38℃ to 48℃. When the setting temperature is changed, the maximum temperature error is about 3℃, and a slight overshoot is occurred. It is about 10%, and this slight overshoot can help to accelerate response time. The time of the adjustment process is about 4-5 minutes, but it is already sufficient to meet the actual control demand. By analysis of the data, the stabilization error is between positive and negative 0.3℃.

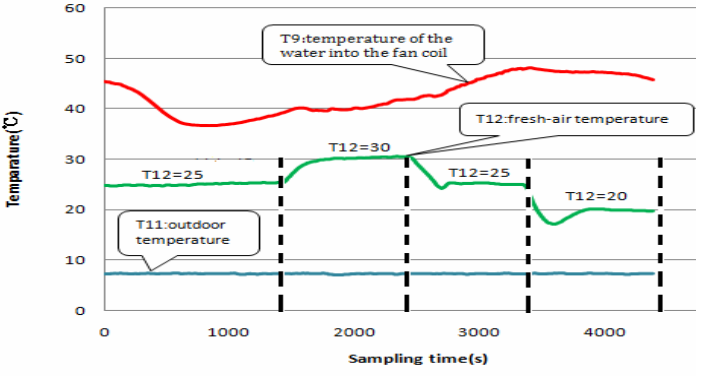


Fig. 6. Different setting temperature control

4.2 The Automatic Transition Process

The principle of the transition process is briefly described as follows:

Firstly, according to the inherent characteristics of this system, it should run in Mode 1 to start solar-collector system to heat fresh air directly when sunshine is sufficient. Secondly, in order to provide enough energy at night, it should run Mode 2 to store energy into PCM tank when the temperature provided by solar energy collection system is 20℃ higher than the set temperature. Lastly, it should run Mode 3 to start PCM tank to heat fresh air directly when sunset or solar energy is not enough.

Taking the transition process of the Mode 2 to Mode 3 as an example, it is easily found from Fig. 7 that the system can work very well in automatic mode transition process. In the Mode 2, the solar energy is not only to heat the fresh air, but also to provide energy for the PCM tank, which can be known from T7, T9 and T12. During the transition process, a little dynamic error is occurred. It is about between positive and negative 3.5℃. After about 4 minutes, the temperature of fresh air stabilizes at the setting temperature within the range of plus or minus 0.3℃, which can be known from the working process under the Mode 3.

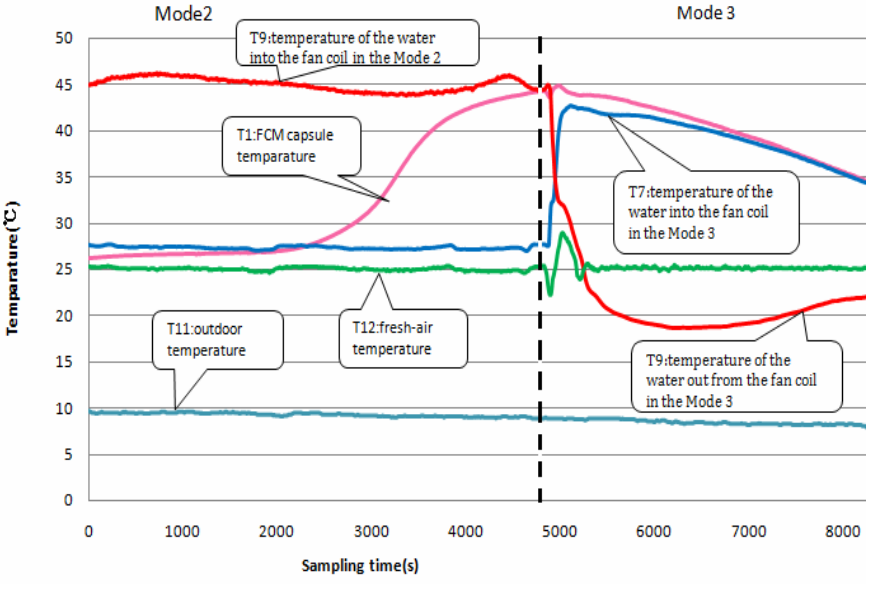


Fig. 7. The automatic transition process

5 Conclusion

In this paper, a fresh air system with PCM tank based on fuzzy control was successfully set up. It works in three different modes. The fuzzy controller design of the three different control rules was completed. The Mode 1 is taken as an example to prove that the system is able to provide fresh air according to different setting temperature. The transition process of the Mode 2 to the Mode 3 is to prove that the

system can adjust automatically according to the change of the solar energy. Through the analysis of the experimental data, it is easy to verify that the system is own high accuracy in stability and able to resist interference. This study will be helpful for the design of fuzzy controller and the applications of solar energy.

References

1. Zhou, L.: Application and Practice of iCAN Field Bus. Guangzhou Zhiyuan Electronics Co., Ltd., Beijing (2007)
2. Shi, J.Z., Gu, F., Goulding, P.: Integration of Multiple Platforms for Real-time Remote Model-based Condition Monitoring. Computers in Industry 58(6), 531–538 (2007)
3. Roonprasang, N.: Experimental Studies of a New Solar Water Heater System Using Solar Water Pump. Energy 33(4), 639–646 (2008)
4. Koca, A., Oztop, H.F., Koyun, T., Varol, Y.: Energy and Exergy Analysis of a Latent Heat Storage System with Phase Change Material for a Solar Collector. Renewable Energy 33(4), 567–574 (2008)
5. Hassan, Marwa, M.: Modeling of An Integrated Solar System. Building and Environment 43(5), 804–810 (2008)
6. Badescu, V.: Optimal Control of Flow in Solar Collector Systems with Fully Mixed Water Storage Tanks. Energy Conversion and Management 49(2), 169–184 (2008)
7. Su, Y., Davidson, J.H.: Discharge of Thermal Storage Tanks via Immersed Baffled Heat Exchangers: Numerical Model of Flow and Temperature Fields. In: Proceedings of the Energy Sustainability Conference 2007, pp. 711–720 (2007)
8. Vikram, D., Kaushik, S., Prashanth, V., Nallusamy, N.: An Improvement in the Solar Water Heating Systems by Thermal Storage Using Phase Change Materials. In: International Solar Energy Conference, Proceedings of the ASME International Solar Energy Conference, SEC 2006, pp. 409–416 (2006)
9. Castell, A., Sole, C., Medrano, M., Nogues, M., Cabeza, L.F.: Comparison of Stratification in a Water Tank and a PCM-water Tank. In: Proceedings of the Energy Sustainability Conference 2007, pp. 465–470 (2007)
10. Qi, Q., Jiang, Y., Ma, Z.: Phase Change Energy Storage Materials and Their Applications in HVAC. Building Enegy& Environment 25(3), 17–19 (2006)

An Efficient Method for Target Extraction of Infrared Images

Ying Ling and Xingjin Mao

School of Computer ScienceNorthwestern Polytechnical University
Shanxi, Xi'an, 710129, China
lybyp@nwpu.edu.cn, Maoxingj1985@gmail.com

Abstract. This paper proposes an efficient method to extract targets from an infrared image. First, the regions of interests (ROIs) which contain the entire targets and a little background region are detected based on the variance weighted information entropy feature. Second, the infrared image is modeled by Gaussian Markov random field, and the ROIs are used as the target regions while the remaining region as the background to perform the initial segmentation. Finally, by searching solution space within the ROIs, the targets are accurately extracted by energy minimization using the iterated condition mode. Because the iterated segmentation results are updated within the ROIs only, this coarse-to-fine extraction method can greatly accelerate the convergence speed and efficiently reduce the interference of background noise. Experimental results of the real infrared images demonstrate that the proposed method can extract single and multiple infrared objects accurately and rapidly.

Keywords: infrared image segmentation; Markov random field; regions of interests; weighted information entropy; iterated condition mode.

1 Introduction

Infrared images provide useful information for many applications, such as military navigation, automatic target recognition and tracking. However, many infrared imaging systems produce images with small signal-to-noise (SNR) and low contrast, which reduces the detectability of targets and impedes further investigation of infrared images. It is a challenging task to extract the targets from an infrared image accurately. Threshold-based method is widely used in infrared target extraction, such as 1-D Otsu [1], 2-D Otsu [2], and max-entropy [3] etc. However, threshold-based methods can not achieve satisfactory results for the infrared images with the complex background and low SNR.

Markov random field (MRF) can capture the spatial constraint among pixels and merge the context information into segmentation process. It gradually develops into a powerful tool for image segmentation.

An infrared target extraction algorithm based on the variance weighted information entropy (WIE) [4] and MRF is proposed in this paper. First, the regions of interest (ROIs) in which all the targets are included are extracted by using the variance WIE.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 185–192, 2010.

Then, the infrared image is modeled with MRF, and the ROIs are used as the target regions and the remaining region as the background to do the initial segmentation. Finally, by searching solution space within the ROIs, the targets are accurately extracted by the energy minimization based on the fact that targets can only exist in ROIs. This coarse-to-fine extraction method can reduce the interference of the background and the noise in a large degree, and decrease the search space greatly.

2 ROIs Detection Based on Variance Weighted Information Entropy

To extract the target coarsely, the ROIs in which all the targets are included are first detected based on the weighted information entropy feature of the image blocks. To evaluate the complex degree of intensity distribution upon an infrared image robustly, the variance WIE [4] is defined as follows.

$$H(\mathrm{s})=-\sum_{S=0}^{255}(\mathrm{s}-\bar{s})^2 * p_s * \log(p_s) \quad . \qquad (1)$$
$$when\ \ p_s = 0, let\ \ p_s * \log(p_s) = 0$$

where s is gray-level $\mathrm{s} \in [0,255]$, p_s is the probability of the grey-level, $\bar{s}$ is the mean intensity of the infrared image.

The variance WIE value of the local region where target locates must be high. On the basis of this, we divide the infrared image into blocks of a certain size, and then the seed block can be obtained according to the variance WIE value of each block. The region growing method is employed on the image blocks to detect a ROI. For the image containing multiple targets, we set the a threshold T_h to detect the multiple ROIs iteratively

$$T_h = k * meanEntropy$$

where $meanEntropy$ is the mean of the variance WIE values of the whole image, and k is a constant typically set to 1~5. For an infrared image I with size $M \times N$, the ROIs detection algorithm is briefly described as follows.

1) Evenly divide the image with size $m \times n$, $m = 2^k, n = 2^l$, $k, l = 1, 2, 3, \ldots$, energy sub-image block is marked with $F(u,v)$, where $0 \le u \le M/m - 1$, $0 \le v \le N/n - 1$.
2) Calculate the variance WIE value $H(u,v)$ for each sub-image $F(u,v)$ using Eq. (1). An entropy image with size of $M/(m-1) \times N/(n-1)$ will be obtained, the mean μ and the standard deviation σ of the variance WIE are calculated synchronously.
3) Obtain seed sub-image. Suppose that the maximal value in the variance WIE image is $H(u,v)$, if $H(u,v) \ge H_T$, where $H_T = \mu + \alpha * \sigma$, α is a constant typically set to 6, and then $F(u,v)$ is seed sub-image. If $H(u,v) < H_T$, let

$m = m/2$, $n = n/2$ and repeat 1) and 2), until the maximal value in the variance weighted information entropy image satisfies $H(u,v) < H_T$ or $m = 2$ or $n = 2$.

4) Judge the existence of target. If $H(u,v) > T_h$, there exists a target in the image, go to Step 5; Otherwise, stop the ROI detection.
5) Detect ROI. On the variance WIE image, starts with the seed sub-image $F(u,v)$, four neighboring pixels are iteratively added to the growing region. The growing result is a rectangular ROI that contains the target.
6) Update image. The grey values of the pixels in the ROI are set to the mean grey value, and then return to Step 1 using the updated image.

3 Accurate Target Extraction Based on MRF

3.1 MAP-MRF Segmentation Framework

Let $S = \{s = (i, j) \mid 1 \le i \le H, 1 \le j \le W\}$ be the set of two-dimensional image lattice sites, where H and W are the image height and width in pixels. X and Y are MRFs defined on neighborhood system η. Y stands for the observed random field and the observed image y is a instance of Y, x is a segmentation configuration of the ideal random field X. The task of image segmentation can be treated as estimating a segmentation configuration x through an instance of the observed field Y. According to the Bayesian rule, the segmentation problem is formulated as:

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x) * P(X = x)}{P(Y = y)} . \tag{2}$$

where $P(Y = y)$ is constant if the observed image is known. $P(X = x \mid Y = y)$ is the posterior probability of $X = x$ conditioned on $Y = y$, and $P(Y = y \mid X = x)$ denotes the probability distribution, which reflects the feature modeling component in segmentation process; $P(X = x)$ is prior probability reflects the region labeling component in segmentation process.

For Markov field, under the guarantee of Hamersley-Clifford [6], we can determine the distribution of $P(X = x)$ via defining potential function in Gibbs distribution. Most MRF based segmentation approaches use the multi-level logistic (MLL) model to define the potential function. Generally, for a segmentation task, the second order pairwise cliques [7] are selected and the potentials of all non-pair wise cliques are defined to be zeros.

$$E_R(x) = \sum_s [\beta \sum_{t \in N_s} \delta(x_s, x_t)] . \tag{3}$$

Here

$$\delta(x_s, x_t) = \begin{cases} -1 & if \quad x_s = x_t \\ 1 & if \quad x_s \neq x_t \end{cases}$$

β is predefined constant. So the form of Gibbs distribution can be denoted as

$$P(X = x) = \frac{1}{Z}\exp(-\frac{1}{T}E_R(x)) \ . \tag{4}$$

where $Z = \sum_{x\in\Omega}\exp[-\frac{1}{T}E_R(x)]$ is named as partition function, T is temperature parameter

MRF model is generally assume that the distribution of all feature data be a Gaussian function [5] with different means and standard deviations. That is

$$P(Y = y \mid X = x) = \prod_s \frac{1}{\sqrt{2\pi}\sigma_m}\exp[-\frac{(y_s - \mu_m)^2}{2\sigma_m^2}] \ . \tag{5}$$

where μ_m , σ_m are the mean and standard deviation for the *m*-th class, that can be estimated by EM algorithm [8].
An energy form of (5) is:

$$E_F = \sum_s [\frac{(y_s - \mu_m)^2}{2\sigma_m^2} + \ln(\sqrt{2\pi}\sigma_m)] \ . \tag{6}$$

So to (2), its energy form is:

$$E = E_R + W * E_F \ . \tag{7}$$

Here E_R denotes the region labeling component, E_F describes the feature modeling component, and W is a weighting parameter to determine how much E_R and E_F individually contribute to the entire energy.

The optimal segmentation $\hat{x}$ could be estimated by minimizing the posterior energy under the MAP criterion

$$\begin{aligned} \hat{x} &= \arg\max(P(X = x \mid Y = y)) \\ &\propto \arg\min E \end{aligned} \ . \tag{8}$$

In [5], Deng proposed a new weighting parameter form which is shown in (9).

$$W = c1 * \alpha^t + c2 \ . \tag{9}$$

where t is the iteration step, α is the drop coefficient. From (9), we can see that at the beginning of iteration the feature modeling component is dominant to capture the feature information. As the iterations go, the region labeling component will interact with the feature modeling component to refine the segmentation result.

3.2 MRF-Based Accurate Target Extraction

The target extraction problem is posed as a binary labeling problem, namely, labeling every pixel i of I with x_i , $x_i \in \{T = 1, B = 0\}$, where T stands for target, and B stands for background . Also we use I^T to denote the ROIs, namely, the target regions, and

I^B ($I^B = I - I^T$) to denote the background region. To model the likelihood of each pixel i belonging to target or background, the target intensity model $P(I_i \mid x_i = 1)$ and the background intensity model $P(I_i \mid x_i = 0)$ are learned from I^T and I^B, respectively.

$$P(I_i \mid x_i = 1) = \frac{1}{\sqrt{2\pi\sigma_T}} \exp\left(-\frac{(I_i - \mu_T)^2}{2\sigma_T^2} \right) . \tag{10}$$

$$P(I_i \mid x_i = 0) = \frac{1}{\sqrt{2\pi\sigma_B}} \exp\left(-\frac{(I_i - \mu_B)^2}{2\sigma_B^2} \right) . \tag{11}$$

where (μ_T, σ_T) and (μ_B, σ_B) denote the mean intensity and standard variance of the target region and the background region, respectively.

After obtaining the ROIs, we know that the targets can only appear in the ROIs. Thus we only need to search the solution within the ROIs. (6) is rewritten as follows.

$$E_F(x_i) = \begin{cases} E_F(x_i = 1) = \infty, \\ E_F(x_i = 1) = -\log(P(I_i \mid x_i = 1)) \\ E_F(x_i = 0) = 0 \qquad\qquad if\ i \in I^B \\ E_F(x_i = 0) = -\log(P(I_i \mid x_i = 0)) \quad if\ i \in I^T \end{cases} . \tag{12}$$

According to (7) and (8), we know that the target segmentation task has been transformed into a combinatorial optimization problem, which is often solved by the simulated annealing (SA) algorithm [9] or the iterated condition mode [10] (ICM). The SA algorithm can find the global maximum in theory. However, it converges slowly, which restricts the application in practice. We use ICM in our work, as it converges fast and the results are satisfying.

In the iterative process, we choose the following function for the variable weighting parameter

$$W = 80 * 0.9^t + 1 . \tag{13}$$

where t is the iteration step.

In summary, our infrared target extraction approach consists of the following steps.

1) Detect ROIs in the infrared image using variance WIE, and obtain the confined areas which contain all the targets.
2) Model the image with the Gaussian MRF.
3) Let the detected ROIs as the target regions, the remaining area as the background to do the initial segmentation.
4) Update the segmentation results using ICM after the initial segmentation, and apply (13) to update the weights during the iterations.
5) If the stopping criterion is met, output the final segmentation result; otherwise go back to step 4 to continue the iteration.

4 Experiments and Results

The proposed target extraction approach is applied to the infrared images with different background, which are sky-sea, ground and sky background. We also compare it with the 2-D Otsu thresholding method and Deng's SMRF method [5] optimized by ICM. All the methods are coded in Matlab 7.0 running on a Pentium4 Q8200 2.33G CPU system with 2G bytes memories.

Fig. 1-4 show targets extraction results of infrared images. (a) of every figure show ROIs detection of original image. (b) of every figure show target extraction result of 2-D Otsu. (c) of every figure show target extraction result of SMRF optimized by ICM. (d) of every figure show target extraction result of the proposed approach. Table 1 compares the running time of the proposed and the SMRF methods. Due to the greatly reduced searching space, the computation time of our method is much less than that of SMRF.

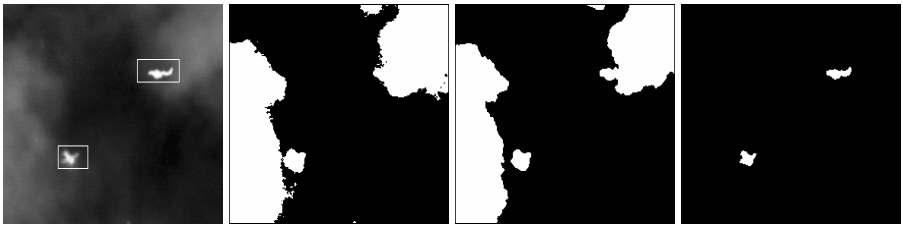
From the results, it is clear that the proposed algorithm is superior to the other two segmentation methods. Because the iterated segmentation results are updated within the ROIs only, this coarse-to-fine extraction method can greatly accelerate the convergence speed and efficiently reduce the interference of the background noise.It can extract single and multiple targets accurately and rapidly.

(a) (b) (c) (d)

Fig. 1. Extraction of multiple infrared targets from sky-sea background using different methods (a) Original images and ROIs detection results; (b) Target extraction results of 2-D Otsu;(c) Target extraction results of SMRF; (d) Target extraction results of the proposed approach.

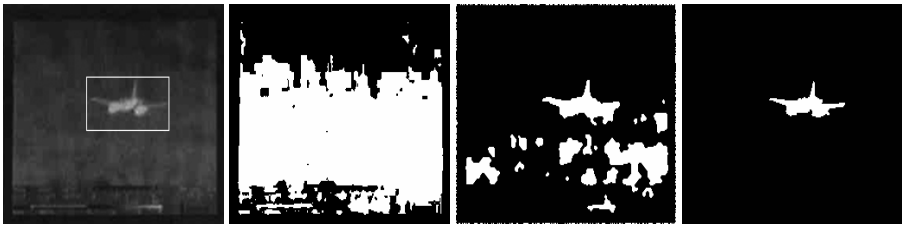
(a) (b) (c) (d)

Fig. 2. Extraction of multiple infrared targets from ground background using different methods. (a) Original image and ROI detection result (b) Extraction result of 2-D Otsu (c) Extraction result of SMRF (d) Extraction result of the proposed approach.



(a) (b) (c) (d)

Fig. 3. Extraction of multiple infrared targets from sky background using different methods. (a) Original image and ROI detection result (b) Extraction result of 2-D Otsu (c) Extraction result of SMRF (d) Extraction result of the proposed approach.



(a) (b) (c) (d)

Fig. 4. Extraction of single infrared target from sky background using different methods. (a) Original image and ROI detection result (b) Extraction result of 2-D Otsu (c) Extraction result of SMRF (d) Extraction result of the proposed approach.

Table 1. Comparison of segmentation time (s)

method	Fig.1	Fig.2	Fig.3	Fig.4
SMRF	408	116	862	1075
Proposed	5.50	1.84	7.51	4.21

5 Conclusions

A coarse-to-fine infrared target extraction algorithm is proposed. The ROIs are first detected by the variance WIE. Then the infrared image is modeled by Gaussian Markov random field using the ROIs as target regions while the remaining region as background to perform the initial segmentation. At last, the targets are accurately extracted by energy minimization using the iterated condition mode by searching solution space within ROIs. Experiments show very promising results for our method.

Acknowledgment

This works was supported by the National Key Laboratory of Fire Control Technology, the National Natural Science Foundation of China (No. 60873086), the

Aeronautical Science Foundation of China (No. 20085153013), and the Fundamental research Found of Northwestern Polytechnical University (No. JC200942).

References

1. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9, 62–66 (1979)
2. Liu, J.Z., Li, W.Q.: The automatic thresholding of gray-level pictures via two-dimensional Otsu method. Acta Autom. Sin (China) 19, 101–105 (1993)
3. Brink, A.D.: Thresholding of digital images using of two-dimensional entropies. Pattern Recognition 25, 803–808 (1992)
4. Yang, L., Zhou, Y., Yang, J.: Variance WIE based infrared images processing. Electron. Lett. 42, 857–859 (2006)
5. Deng, H.W., Clausi, D.A.: Unsupervised image segmentation using a simple MRF model with a new implementation scheme. Pattern Recognition 37, 2323–2335 (2004)
6. Li, S.Z.: Markov Random Field Modeling in Computer Vision. Springer, Tokyo (1995)
7. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. 6, 721–741 (1984)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. B-1, 1–38 (1977)
9. Metopolis, N.: Equations of state calculations by fast computational machine. Journal of Chemical Phusics 21, 1087–1097 (1953)
10. Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society 48, 259–302 (1986)

Second Order Central Difference Filtering Algorithm for SINS/GPS Integrated Navigation in Guided Munitions

Lei Cai[1,3] and Xuexia Zhang[2]

[1] College of Control Science and Engineering,Shangdong University, Jinan 250061, China
[2] College of Animal Science,Henan Institute of Science and Technology, Xinxiang 453003, China
[3] Troops 94569 PLA, Jinan 250023, China
cailei1998@sohu.com

Abstract. With the SINS/GPS integrated navigation system, the guided munitions can be carried out in complex weather, and have the high positioning navigation accuracy. This paper deduces key matrix of the second order central difference filtering (CDF2) equation. The CDF2 is described as a sigma point filter in a unified way where the filter linearizes the nonlinear dynamic and measurement functions by using an interpolation formula through systematically chosen sigma points. The effect which the key parameters of CDF2 bring to information fusion is analyzed qualitatively. The structure of loose integration is also given. According to the test data, the fusion algorithm based on CDF2 is applied. Compared to the original algorithm in longitude, latitude, altitude and velocity, the orientation precision is improved greatly.

Keywords: SINS; GPS; integraion navigation system; information fusion.

1 Introduction

A certain type of the guided bombs is an air-to-ground weapons which can have the performance of being sent out in the high-altitude and long-distance condition and can work in all-weather environments. It comes from the common bombs installed with the flight guidance system and has the performances of the low cost and great effectiveness. With the SINS/GPS integrated navigation system, it can be carried out in complex weather, and have the high positioning navigation accuracy.

However, there are several situations where GPS experience either total system outage (due to satellite signal blockage) or deterioration of accuracy (due to multipath effects and clock bias error). Therefore, GPS is usually combined with Inertial Navigation System (INS), which is a self-contained system incorporating three orthogonal accelerometers and three orthogonal gyroscopes. These monitor the guided munitions' linear accelerations and rotation rates[1]. A set of mathematical transformations and integrations with respect to time are applied to these raw measurements to determine position, velocity and attitude information. However, the INS accuracy deteriorates with time due to possible inherent sensor errors (white

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 193–200, 2010.

noise, correlated random noise, bias instability, and angle random walk) that exhibit considerable long-term growth.

The central difference filter (CDF2), also called divided difference filter (DDF), is similar to the concept behind the UKF, both of which are considered to be among the class of linear regression Kalman filters, but is based on a multivariable extension of Stirling's interpolation formula. Like the UKF, the CDF2 generates several points about the mean based on varying the covariance matrix along each dimension. A slight difference from the UKF is in that the CDF2 evaluates a nonlinear function at two different points for each dimension of the state vector that are divided proportionally by the chosen step size [6].

In this research, we aim at developing an CDF2-based GPS/INS integration module taking into consideration the INS error trend and thus providing a better positioning accuracy during both short and long GPS outages. Such technique combines the advantages of some of the existing models with the advantages of the central difference filter in representing the sequential process in the input data (INS position or velocity). In this way, it should be possible for the proposed module to model both the INS position and velocity errors based on the current and some past samples of the INS position and velocity, respectively.

2 CDF2 Filtering Method

2.1 Problems in the Conventional EKF

The nonlinear system's dynamic and observation equations in discrete form are given as

$$x_k = f\left(x_{k-1}, k-1\right) + \omega_{k-1} \tag{1}$$

$$z_k = h_k\left(x_k\right) + \vartheta_k \tag{2}$$

where $f_k(\cdot) \in R^{n\times n}$ is the process model; $x_k \in R^n$ is the state vector; $\omega_k \in R^n$ is the system noise; $Z_k \in R^P$ is the measurement vector; $h_k \in R^{p\times n}$ is the measurement model and $\vartheta_k \in R^p$ is the measurement noise.

It is assumed that noise is uncorrelated Gaussian white noise sequences with mean and covariances given as

$$E\{\omega_i\} = 0\,; E\{\omega_i \omega_j\} = Q\delta_{ij}$$

$$E\{\vartheta_i\} = 0\,; E\{\vartheta_i \vartheta_j\} = R\delta_{ij}$$

$$E\{\omega_i \vartheta_j\} = 0$$

where $E\{\cdot\}$ denotes the expectation, and δ_{ij} is the Kronecker delta function. Q and R are bounded positive definite matrices $(Q > 0, R > 0)$. Initial state x_0 is normally distributed with zero mean and covariance P_0.

2.2 CDF2 Filtering Method

This is an exciting development in the derivative-free nonlinear state estimation technique. The CDF2 is described as a sigma point filter (SPF) in a unified way where the filter linearizes the nonlinear dynamicand measurement functions by using an interpolation formula through systematically chosen sigma points.

Conceptually, the realization principle resembles that of the EKF; however, it is significantly simpler because it is not necessary to formulate the Jacobian matrices of the nonlinear dynamic and measurement equations. Thus, the CDF2 can replace the EKF, and its higher-order estimators in practical real-time applications that require accurate estimation, but less computational cost. The CDF2 filter makes use of first and second order central differences (CDs) to approximate nonlinear transformation of the state and covariance. The CDF2 algorithm is implemented in accordance with the block diagram shown in Fig. 1.

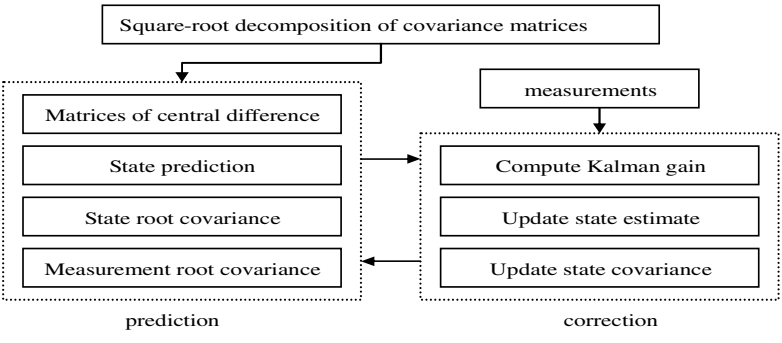


Fig. 1. The central difference filter

3 SINS Nonlinear Error Modeling

The basis of the error analysis is, of course, the error model that describes the propagation of the errors. The model, besides being the tool for the error analysis, also plays an important role in the implementation of integration filters and in the real time failure detection.

The conventional linearized error models of SINS are found not to be effective to represent the nonlinear characteristics of navigation error propagation in case of the system which causes the large navigation errors. To represent the large azimuth error of the SINS, the axes misalignment angles are introduced and the relationships for the velocity, position and attitude errors are derived.

3.1 Coordinate Frames

Coordinate frames used in this study of integrated navigation system are defined.

Inertial frame (i-frame): It has origin at Earth's center; z-axis is normal to the equatorial plane; x-axis lies in equatorial plane, its direction can be specified arbitrarily; y-axis complements the right-handed system.

Body frame (b-frame): It has origin at center of mass of vehicle; x-axis is along longitudinal axis; z-axis is perpendicular to longitudinal plan of symmetry and yaxis complements the right-handed system.

The Earth fixed frame (e-frame): It is the Earth fixed coordinate frame used for position location definition. Its z-axis is coincident with the Earth's polar axis while the other two axes are fixed to the Earth within the equatorial plane.

The navigation frame (n-frame): It is a local geographic coordinate frame; z-axis is parallel to the upward vertical at the local Earth surface referenced position location; x-axis points towards east, and y-axis points toward north. Error analysis schema shown in Fig. 2 portrays the relationship between different frames.

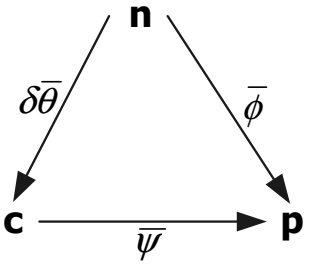


Fig. 2. Error analysis schema

Here, c and p refer to computed and mathematical platform frames, respectively. The p frame has its origin at the IMU computed position and the IMU output defines its orientation.

The c frame is a perfectly local level frame with origin at the IMU computed position. It is obvious that c follows n . p follows c and p follows n indirectly. The misalignments between the n and p , p and c , c and n are expressed by small angles $\bar{\phi}$, $\bar{\psi}$ and $\delta\bar{\theta}$ respectively. It follows that

$$\bar{\phi} = \delta\bar{\theta} + \bar{\psi} \tag{3}$$

3.2 Attitude Error Model

The attitude error for the local level SINS mechanization is the orthogonal transformation error between the body and computed frame. In this perspective, nonlinear psi angle attitude error model of the SINS with large heading error and small tilt errors is given by

$$\dot{\bar{\psi}} = \begin{bmatrix} 1-\cos\psi_z & -\sin\psi_z & \psi_y \\ \sin\psi_z & 1-\cos\psi_z & -\psi_x \\ -\psi_x\sin\psi_z - \psi_y\cos\psi_z & \psi_x\cos\psi_z - \psi_y\sin\psi_z & 0 \end{bmatrix} \times \left(\bar{\omega}_{ie}^{f} + \bar{\omega}_{ec}^{f}\right) + \bar{\omega}^{p} \tag{4}$$

where v is the velocity; ψ represents attitude error; φ is geodetic latitude; ω_{ie} is the earth's rate; r_λ is the normal radius of curvature for east-west motion; r_φ is the meridianal radius of curvature for north-south motion; ε is the gyros drift, and p and c in the superscript stands for mathematical platform frame and computer frame, respectively.

3.3 Velocity Error Model

The velocity error model with large heading error and small tilt errors is

$$\dot{\delta v}^c = \begin{bmatrix} 1-\cos\psi_z & \sin\psi_z & -\psi_x\sin\psi_z - \psi_y\cos\psi_z \\ -\sin\psi_z & 1-\cos\psi_z & \psi_x\cos\psi_z - \psi_y\sin\psi_z \\ \psi_y & -\psi_x & 0 \end{bmatrix} \times \bar{f}^p - \left(2\bar{\omega}_{ie}^c + \bar{\omega}_{ec}^c\right)\times\delta\bar{v}^c + \delta g^c + \bar{\nabla}^p \tag{5}$$

where ∇ stand for accelerometer's bias; f is the specific force, and δg is the acceleration error due to gravity.

3.4 Position Error Model

The position error model expressed in computer frame is given as

$$\left[\delta\dot{\varphi} \quad \delta\dot{\lambda} \quad \delta\dot{h}\right]^T = \left[\frac{\delta v_y^c}{r_\varphi + h} \quad \frac{\sec\varphi\delta v_x^c}{r_\lambda + h} + \frac{v_x^c\sec\varphi\tan\varphi\delta\varphi}{r_\lambda + h} \quad v_z^c\right]^T \tag{6}$$

where $\delta\varphi$ and $\delta\lambda$ denotes latitude and longitude errors, respectively, and δh is height error.

3.5 Attitude Correction

The corrected attitude matrix is obtained as

$$C_b^n = C_p^n C_b^p = \left(I + \bar{\phi}\times\right)C_b^p \tag{7}$$

where C represents the transformation matrix between subscripted frame to the superscripted frame, $\bar{\phi}\times$ is the skew symmetric matrix of attitude error angles, and I is the identity matrix.

At this point, $\bar{\phi}$ is estimated using (3) where the error angles $\bar{\psi}$ are estimated by the integration filter and $\delta\bar{\theta}$ are acquired from the position error states as

$$\delta\bar{\theta} = \left[-\frac{\delta r_y^n}{r_\varphi + h} \quad \frac{\delta r_x^n}{r_\lambda + h} \quad \frac{\delta r_x^n\tan\varphi}{r_\lambda + h}\right]^T \tag{8}$$

where $\delta\bar{r}$ represents position errors in rectangular coordinate system acquired via integrating the velocity errors as estimated by the integration filter.

4 Experimentation

The SINS comprises a micro-electro-mechanical system (MEMS) based low cost strapdown inertial measurement unit (SIMU) mounted on a land vehicle. It provides increments in attitude and velocity for navigation at 10 Hz. The SIMU contains three vibrating quartz accelerometers having bias in-run stability from turn-on to turn-on equal to 0.5 milli-g. Attitude is sensed by three quartz attitude sensors having drift in-run stability from turn-on to turn-on equal to 4 deg/hour. The GPS Jupiter LP receiver has twelve channel course/Acquisition (C/A) code, L1 frequency engine. The GPS data employed is available at 1 Hz.

The discrete filter realization used in this paper is in the direct feedback mode where the estimated attitude errors are fed back to the SINS, thus minimizing the evolution of the observed velocity errors those are to be delivered as an observation to the filter. In this algorithm, quaternion is obtained from the corrected attitude matrix and is fed back for attitude error compensation.

Comparison of CDF2 and EKF in orientation precision is shown in Fig. 3 to Fig.6.

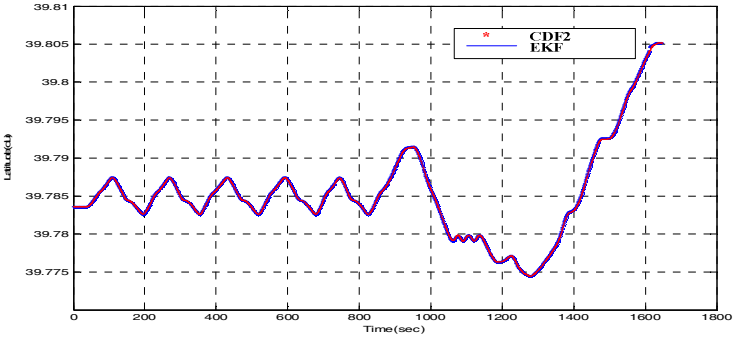


Fig. 3. Latitude component

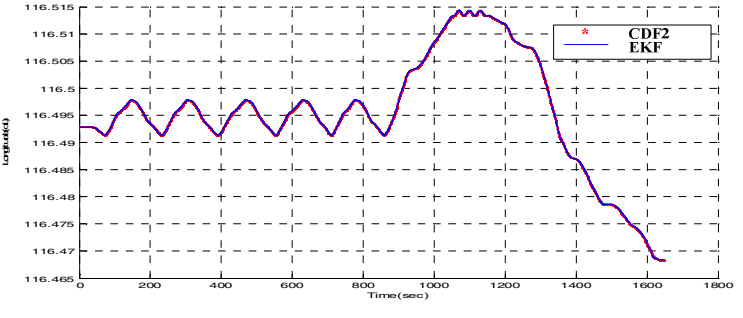


Fig. 4. Longitude component

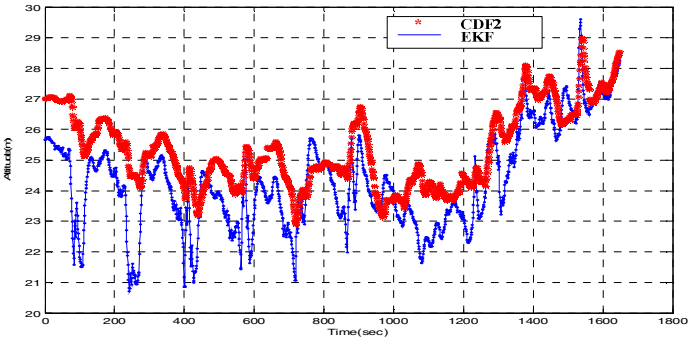


Fig. 5. Altitude component

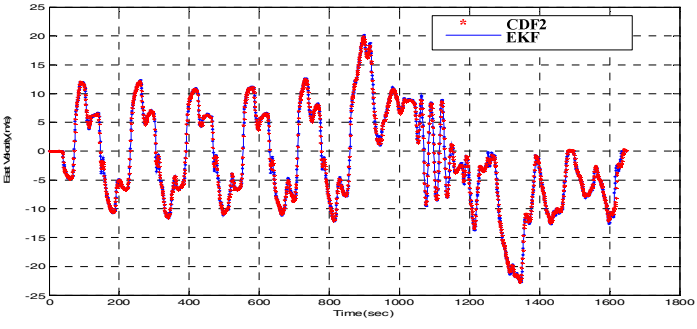


Fig. 6. Longitude component

The experimental results depicting performance comparison among the two integration filters are presented. Figure 3 presents position and its RMSE assessment. Test results shown illustrate comparison of velocities and their RMSE. A magnified portion of the results is shown to corroborate the performance of the applied nonlinear data synthesis filters.

As can be seen from the results, while the satellites disappear below the visibility mask, the EKF can no longer keep track of the INS drifts, given that it can not appropriately make use of the dynamic absorbability of the system, deteriorated by linearization. This is for the reason that nonlinearities in the Jacobin most likely do not approximate the functions well, especially for values farther from the nominal point around which the function was linearized. In usual circumstances, when enough number of satellites is visible, the EKF linearization may be deemed a good approximation of the problem. From the results, it is evident that there is insignificant difference in the performance of the UKF and the second order CDF, and both of these filters are anticipated to demonstrate comparable accuracy.

References

1. Saulson, B.G.: Nonlinear estimation comparison for ballistic missile tracking. Automatica 43, 1424–1438 (2004)

2. Schei, T.S.: A finite-difference method for linearization in nonlinear estimation algorithms. Automatica 51, 252–260 (2003)
3. Kotecha, J.H., Djuric, P.M.: Gaussian particle filtering. IEEE Transactions on Signal Processing 51, 2592–2601 (2003)
4. Noureldin, A., El-Shafie, A., Taha, M.R.: Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation. Eng. Appl. Artif. Intell. 20, 49–61 (2007)
5. Noureldin, A., Osman, A., El-Sheimy, N.: A neuro-wavelet method for multisensor system integration for vehicular navigation. J. Meas. Sci. Technol. 15, 404–412 (2004)
6. Semeniuk, L., Noureldin, A.: Bridging GPS outages using neural network estimates of INS position and velocity errors. Meas. Sci. Technol. 17(9), 2783–2798 (2006)

Adaptive Signal Processing for ARX System Disturbed by Complex Noise*

Yulai Zhang and Guiming Luo

School of Software, Tsinghua University
Beijing 100084, China
gluo@mail.tsinghua.edu.cn

Abstract. The inverse of the Fisher information matrix can be decided by the system input sequence and the disturbance variance if a Gaussian noise is involved. The lower bound mean-square error matrix of any unbiased estimator is given by Cramer-Rao Lemma. When a system is disturbed by some biased noises, the classical Fisher information matrix would be not valid. The bound is not fitted when a biased estimator is implemented. Signal processing for ARX model disturbed by complex noise is concerned in this paper. Cramer-Rao bound of a biased estimation is obtained. An adaptive signal processing algorithm for identification of ARX system disturbed by biased estimation is proposed. Some experiments are included to verify the efficiency of the new algorithm.

Keywords: Cramer-Rao bound; adaptive signal processing; LMI-s; biased estimator.

1 Introduction

In recent years, with the development of signal processing and modeling technique, obtaining an accurate model by using minimum input energy becomes an attractive topic [1] [2] in many application fields. The quality of an estimator can be assessed by the mean-square error matrix of its parameter vector. For an unbiased estimator, the lower limit to the value of this matrix can be obtained by the Cramer-Rao inequality.

An iterative identification method based on optimal input design has been outlined in Lindqvist and Hjalmarsson [3]. In each step of iteration a semi-definite optimal problem is solved such that required parameter accuracy will be achieved under a minimum input energy. In Gerencser and Hjalmarsson [4] a recursive improvement on the previous algorithm is given and the convergence property is also analyzed. This recursive identification method is restricted to a smaller class of the problem, i.e. the identification of ARX models with FIR input filter. Recursive Least-Squares method (RLS) is used to identify parameters in ARX model under

* This work is supported by the National Natural Science Foundation of China No.60973049 and No.60635020.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 201–208, 2010.

Gaussian white noise as the system interruption. This makes the estimation to be unbiased. In this kind of algorithm Cramer-Rao bound [5] is used to refine its parameter accuracy.

A basic assumption of Cramer-Rao bound is the unbiased parameter estimator [6]. If parameter estimation is biased, which often occurs in the practice, the Cramer-Rao equality will not be strict enough, and a more precise bound should be concerned and analyzed.

In this paper, the Cramer-Rao inequality under an assumption of biased parameter estimate is given in the first part. Then this new inequality is used to construct an optimal input design algorithm for ARX model with a biased noise as the system interruptions. Then in section 4 some algorithm details and the simulation results are presented.

2 Modified Cramer-Rao Inequality

The Cramer-Rao inequality [5] is proofed under the assumption that the estimation of system parameters is unbiased. This inequality will not be accurate enough in circumstances of biased parameter estimation.

The assumption of the biased parameter estimation can be written as

$$E\hat{\theta} = \theta_0 + \xi(\theta_0, x^N) \tag{1}$$

$$\theta_0 + \xi(\theta_0, x^N) = \int_{R_N} \hat{\theta}(x^N) f_y(\theta_0, x^N) dx^N \tag{2}$$

where θ_0 denotes the system parameter vector and $\hat{\theta}$ is its estimated value, x^N denotes the input sequence and f_y is the probability density function of y^N, which is the output sequence [6]. Differentiate (2) with respect to θ_0 gives

$$\begin{aligned} I + \frac{d\xi(\theta_0, x^N)}{d\theta_0} &= \int_{R_N} \hat{\theta}(x^N)[\frac{d}{d\theta_0} f_y(\theta_0, x^N)]^T dx^N \\ &= \int_{R_N} \hat{\theta}(x^N)[\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)]^T \cdot f_y(\theta_0, x^N) dx^N \\ &= E\left(\hat{\theta}(y^N)[\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)]^T \right) \end{aligned} \tag{3}$$

Here I is the unit matrix and R^N the Euclidian N-dimensional space. By definition we have

$$1 = \int_{R^N} f_y(\theta_0, x^N) dx^N \tag{4}$$

Differentiate the above expression with respect to θ_0 either gives

$$\begin{aligned} 0 &= \int_{R_N} [\frac{d}{d\theta_0} f_y(\theta_0, x^N)]^T dx^N \\ &= \int_{R_N} [\frac{d}{d\theta_0} \log f_y(\theta_0, x^N)]^T f_y(\theta_0, x^N) dx^N \\ &= E[\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)]^T \end{aligned} \tag{5}$$

Expectations in (3) and (5) are with respect to y^N . Multiple (5) by θ_0 and subtract it from (3)

$$E[\hat{\theta}(y^N) - \theta_0][\frac{d}{d\theta_0} \log f_y(\theta_0, y^N)]^T = I + \frac{d\xi(\theta_0, x^N)}{d\theta_0} \tag{6}$$

Denote

$$\alpha = \hat{\theta}(y^N) - \theta_0, \beta = \frac{d}{d\theta_0} \log f_y(\theta_0, y^N), X = I + \frac{d\xi(\theta_0, x^N)}{d\theta_0}$$

So (6) can be rewritten as

$$E\alpha\beta^T = I + \frac{d\xi(\theta_0, x^N)}{d\theta_0} \tag{7}$$

Hence

$$E\begin{bmatrix} \alpha \\ \beta \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}^T = \begin{bmatrix} E\alpha\alpha^T & X \\ X^T & E\beta\beta^T \end{bmatrix} \geq 0$$

Using Lemma given in [6], we have

$$E\alpha\alpha^T \geq X(E\beta\beta^T)^{-1}X^T \tag{8}$$

Denote M as the original Fisher Information Matrix, than the new inequality should be

$$\begin{aligned} &E[(\hat{\theta}(y^N) - \theta_0)(\hat{\theta}(y^N) - \theta_0)^T] \\ &\geq (I + \frac{d\xi(\theta_0, x^N)}{d\theta_0}) M^{-1} (I + \frac{d\xi(\theta_0, x^N)}{d\theta_0})^T \end{aligned} \tag{9}$$

It is clear that the difference of the biased estimation Cramer-Rao inequality depends on the differentiate matrix of the parameter estimate bias. If the estimation bias is 0 or θ_0 is not an argument of the estimate bias ξ, the inequality will turn into its original format.

3 Adaptive Input Signal Algorithm

3.1 The System

Consider system (10) with a specific non-white noise in (11)

$$A^*(q^{-1})y(k) = B^*(q^{-1})u(k) + e(k) \tag{10}$$

$$e(k) = A^*(q^{-1})w(k) \tag{11}$$

where $A^*(q^{-1})$ and $B^*(q^{-1})$ are polynomials with degree n_a and n_b, q^{-1} stands for the backward shift operator

$$A^*(q^{-1}) = 1 + \sum_{i=1}^{n_a} a_i^* q^{-i}, B^*(q^{-1}) = \sum_{i=1}^{n_b} b_i^* q^{-i}$$

It is assumed that $A^*(q^{-1})$ is stable. For the interruption part, $w(k)$ is a Gaussian white noise $E[w(k)] = 0, E[w^2(k)] = \sigma_w^2$, so $e(k)$ is colored by filter A.

Since the RLS identification method has the same convergence property with its non-recursive counterpart. So we look into the Least-Square (LS) method to get the differentiate matrix of the parameter estimate bias. Under non-white noise in (11), the LS algorithm gives a biased estimation, $\varphi(k)$ is the regression vector at time k,

$$E(\hat{\theta}_{LS} - \theta_0) = -\sigma_w^2 R^{-1} H \theta_0 \tag{12}$$

$$R = \lim_{L\to\infty} \frac{1}{L} \sum_{k=1}^{L} \varphi(k)\varphi^T(k)$$

$$H = \begin{bmatrix} I_{n_a} & 0 \\ 0 & 0_{n_b} \end{bmatrix}$$

3.2 Adaptive Input Design Algorithm

It is show in [4], for ARX models, we can formulate the adaptive input design problem as the follow semi-definite programs

$$\begin{aligned} &\min_{r,Q} r_0 \\ &s.t. Var(\| \hat{G}_N \|_2^2) < \gamma \\ &\Phi_u(e^{j\omega}, r) \geq 0 \end{aligned} \tag{13}$$

where Q is a auxiliary matrix for the positive of input spectrum and

$$r=[r_0,...,r_m], r_\tau = E[u_n u_{n-\tau}], \Phi_u(e^{j\omega}, r) = \sum_{\tau=-m}^{m} r_{|\tau|} e^{j\omega\tau}$$

Here $Var(\| \hat{G}_N \|_2^2)$ is defined as the variance of the following parameter expression

$$\| G^* \|_2^2 = (\theta^*)^T \theta^*$$

and γ is the anticipated accuracy. Positive-real Lemma can be used here to keep the input spectrum positive. Let

$$A=\begin{pmatrix} 0 & 0 \\ I_{m-1} & 0 \end{pmatrix}, B=\begin{bmatrix} 1 & 0 & ... & 0 \end{bmatrix}^T, C=[r_1 r_2 \cdots r_m], D=\frac{1}{2} r_0$$

So the positive of input spectrum is equivalent to the existence of a symmetric matrix $Q \geq 0$ such that

$$\begin{bmatrix} Q-A^TQA & -A^TQB \\ -B^TQA & -B^TQB \end{bmatrix} + \begin{bmatrix} 0 & C^T \\ C & 2D \end{bmatrix} \geq 0 \tag{14}$$

Than the variance constraint can be re replaced by its approximation:

$$4\theta^*(I+\sigma_w^2 R^{-1}H)^T R^{-1}(I+\sigma_w^2 R^{-1}H)\theta^* \leq \gamma$$

Using Schur complements the above constraint can be formulated as an LMI in variable r as follow

$$\begin{pmatrix} (I+\sigma^2 R^{-1}H)^{-1} R (I+\sigma^2 R^{-1}H)^{-T} & 2\theta^* \\ 2\theta^* & \frac{\gamma N}{\sigma_w^2} \end{pmatrix} > 0 \tag{15}$$

Notice that matrix R contains r as its argument.

True values of the unknown system are involved in this expression and this makes it impossible for calculating. So we replace it by the estimate value in each step. And the estimate values are given by a standard RLS algorithm.

Acceptable initial values should be given by algorithm users. Other details of construction as well as resolution of such a LMI problem will be given in section 4 by an explicit example. From (14) and (15), we can rewrite (13)

$$\begin{aligned} &\min_{r,Q} r_0 \\ &s.t. \begin{pmatrix} (I+\sigma_n{}^2 R^{-1}H)^{-1} R (I+\sigma_n{}^2 R^{-1}H)^{-T} & 2\hat{\theta}_n \\ 2\hat{\theta}_n & \frac{\gamma N}{\sigma_n^2} \end{pmatrix} > 0 \\ &\begin{bmatrix} Q-A^TQA & -A^TQB \\ -B^TQA & -B^TQB \end{bmatrix} + \begin{bmatrix} 0 & C^T \\ C & 2D \end{bmatrix} \geq 0 \end{aligned} \tag{16}$$

Each recursive step of RLS gives a new estimation of system parameters. The forward input signal filter should also be renewed by implementing (16) as well as the filter design method in section 3.3 for each time. We notice that the input sequence is non-stationary.

3.3 Forward Filter Design

A map from r to forward filter F is defined in Anderson and Moore [7] and [8]. The designed system input signal will be the output of this filter F with a Gaussian white noise input to F. And notice that this input to filter F is independent from the system disturbance.

4 Numerical Experiments

4.1 Modified Cramer-Rao Bound Comparison

This comparison simulation is done on standard RLS algorithm without forward FIR filter design in order to eliminate the influence of non-stationary input. Consider system,

$$y(t) - ay(t-1) = bu(t-1) + w(t) - aw(t-1) \tag{17}$$

Set $\theta^* = [a,b] = [0.4, 2.5]$, the initial values of RLS are $\theta = [0.01, 0.01]$, $P = 1000 * I$, $\sigma^2 = 1 \times 10^{-6}$, recursive step number N=200. And a Gaussian white noise is given as the input sequence u which $E(u_k^2) = 0.1$ and $\sigma_w{}^2 = 1$ as the system disturbance.

The following table shows the minimum Eigen value of the covariance matrix and its bound computed by both Cramer-Rao inequalities. Results are different from time to time, but the new Cramer-Rao inequality gives a more precise bound in all simulations.

Table 1. Comparison of Unbiased and Biased Cramer-Rao Bound

	minimum Eigen value
Unbiased Cramer-Rao Bound	0.0029
Biased Cramer-Rao Bound	0.0071
Parameter covariance matrix	0.0288

4.2 Simulation Results

Matrix R in (16) should be expressed as R(r), where r is the unknown spectrum vector. The optimization problem is dealt with Matlab toolbox YALMIP [9] and its solver sdpt3. Since there is no guarantee to what the RLS estimator may give at the beginning of the procedure, it is necessary to project estimated parameters into a set K so that the estimated model will be stable. Otherwise there will be no solution for (16) [4][10]. For the system structure (17), we can simple restrict first parameter of θ in (-1, 1).

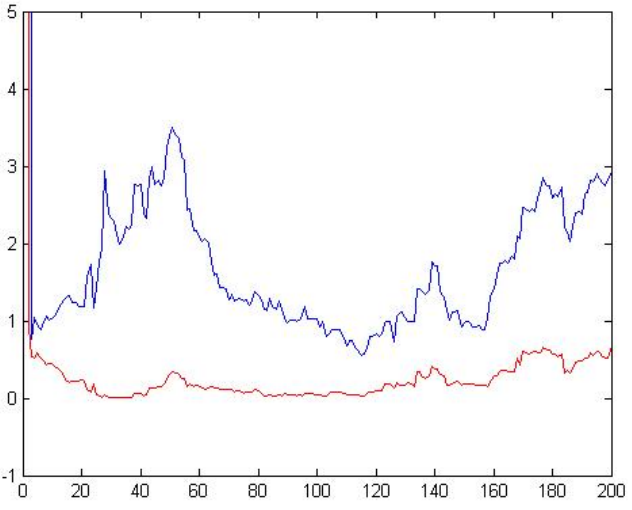


Fig. 1. Parameter Estimation Error

Simulation results are presented in Figure 1. We use system (17) and the RLS initial values are the same with section 4.1.

Figure 1 shows the parameter estimation error of the algorithms in this paper and its counterpart in [4] with the same realization of system interruption sequence. Red Line shows system parameters error estimated by algorithm in this paper. Blue Line shows system parameters accuracy estimated by algorithm in [4]. The algorithm depict in blue line does not perform to achieve its target under biased parameter estimation in 200 steps. The algorithm in red line gives a better performance under the same situation.

5 Conclusions

In this paper, Cramer-Rao inequality under biased parameter estimation is given and the existing optimal input design algorithm based on forward filter design is modified to fit the biased RLS algorithm for ARX models. We proofed that the new equality gives a more precise bound for biased estimation structure, and consequently, a better performance is observed when implement this new bound in the adaptive input signal design algorithm. Finally, some simulation examples are given to demonstrate the new bound and the new algorithm.

References

1. Goodwin, G., Payne, R.L.: Dynamic system identification: Experiment design and data analysis. Academic Press, New York (1977)
2. Bombois, X., Scorletti, G., Gevers, M., Hildebrand, R., Van den Hof, P.: Cheapest open-loop identification for control. In: Conference on Decision and Control. IEEE, Paradise Island (2004)

3. Lindqvist, K., Hjalmarsson, H.: Identification for control: Adaptive input design using convex optimization. In: Proc. 40th IEEE Conf. on Decision and Control (2001)
4. Gerencser, L., Hjalmarsson, H., Martensson, J.: Identification of ARX systems with non-stationary inputs – asymptotic analysis with application to adaptive input design. Automatica 45, 623–633 (2009)
5. Cramer, H.: Mathematical methods of Statistics. Princeton University Press, Princeton (1946)
6. Ljung, L.: System identification: Theory for the user, 2nd edn. Prentice Hall, Englewood Cliffs (1999)
7. Anderson, B.D.O., Moore, J.B.: Optimal filtering. Prentice-Hall, Englewood Cliffs (1979)
8. Stoorvogel, A.A., Saberi, A.: The discrete algebraic riccati equation and linear matrix inequality. Linear Algebra and its Application, 274–365 (1998)
9. Lofberg, J.: YALMIP: A toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference (2004)
10. Gerencser, L., Hjalmarsson, H.: Adaptive input design in system identification. In: Proc. of the 44th IEEE Conference on Decision and Control (2005)

Abstraction for Model Checking the Probabilistic Temporal Logic of Knowledge*

Conghua Zhou, Bo Sun, and Zhifeng Liu

School of Computer Science and Telecommunication Engineering, Jiangsu University
Xuefu Road No. 301, Zhenjiang, Jiangsu Province 212013, China
chzhou@ujs.edu.cn

Abstract. Probabilistic temporal logics of knowledge have been used to specify multi-agent systems. In this paper, we introduce a probabilistic temporal logic of knowledge called PTLK for expressing time, knowledge, and probability in multi-agent systems. Then, in order to overcome the state explosion in model checking PTLK we propose an abstraction procedure for model checking PTLK. The rough idea of the abstraction approach is to partition the state space into several equivalence classes which consist of the set of abstract states. The probability distribution between abstract states is defined as an interval for computing the approximation of the concrete system. Finally, the model checking algorithm in PTLK is developed.

Keywords: agent, model checking, abstraction, probabilistic logic, temporal logic, epistemic logic.

1 Introduction

Model checking[1] is a powerful technique for verifying systems and detecting errors at early stages of the design process, which is obtaining wide acceptance in industrial setting. The state explosion problem is the major obstacle in making model checking tractable. Therefore, there exists much work focusing on this famous problem, and several techniques have been suggested to overcome this limitation of model checking. In all these techniques, abstraction[2, 3, 4] is one of the most successful techniques for overcoming the state space explosion problem in model checking. This motivates us to use the abstraction to reduce the state space in model checking probabilistic temporal logics of knowledge.

A range of logics combining temporal, epistemic, and probabilistic components such as PFKD45[5], PET L[6] have been developed for specifying randomized or probabilistic programs, communication protocols, and complex systems such as distributed or multi-agent systems. In these logics, probabilistic components are only

* Supported by the National Natural Science Foundation of China No. 60773049, the People with Ability Foundation of Jiangsu University No. 07JDG014, the Fundamental Research Project of the Natural Science in Colleges of Jiangsu Province No. 08KJD520015, the Ph.D. Programs Foundation of Ministry of Education of China No. 20093227110005.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 209–221, 2010.

used to reason about the uncertainty of knowledge. It is natural to extend these logics to consider probabilistic aspects of system such as "consensus will eventually be reached with probability 1", "with 99% chance the system will not fail". So, in this paper we introduce the probabilistic component into temporal epistemic logic CTLK[7, 8, 9] yielding a new probabilistic temporal logic of knowledge called *PTLK*. Then a probabilistic Kripke structure is given for interpreting *PTLK*, and the model checking algorithm based on the algorithm for model checking CTL is presented, which given a structure M and a *PTLK* formula ϕ determines whether the structure M satisfies ϕ.

In this paper our most important work is to simplify a system, while preserving some of its relevant behaviors, so that the simplified system is more accessible to model checking procedure and is still sufficient to establish certain properties. Let M be the concrete system that we want to verify against some probabilistic temporal epistemic formula. The rough idea of the abstraction approach is to partition the state space of M into several equivalence classes which consist of the set of abstract states. The probability distribution between abstract states is an interval. For different kinds of formulas, we can choose the corresponding probability function over the interval such that the abstract model A satisfying ϕ implies M satisfies ϕ. To our knowledge, this is the first attempt to provide abstraction techniques that are useful to prove properties expressed in the probabilistic temporal logic of knowledge.

The rest of this paper is organized as follows. In Section 2, we present the basic definition of the probabilistic temporal logic of knowledge *PTLK* and the probabilistic Kripke structure. In Section 3, we present the principle of abstracting the probabilistic Kripke structure, and show that the abstraction preserves some *PTLK* formula. The model checking algorithm for *PTLK* is presented in Section 4. A case of the dining cryptographers protocol is studied in Section 5. Conclusions and future work are summarized in Section 6.

2 Probabilistic Temporal Logic of Knowledge

In this section, we define a logic, called Probabilistic Temporal Logic of Knowledge(*PTLK*), for expressing time, knowledge, and probability in multi-agent systems. Assume *Ap* is a finite set of atomic propositions, $Ag = \{1,\ldots, n\}$ is a finite set of agents. Formulas in *PTLK* are built from atomic propositions, propositional logic connectives, temporal operators, and knowledge operators for expressing time and probability.

Definition 1 (PTLK Syntax). The syntax of *PTLK* formulas over *Ap* and *Ag* is defined inductively as follows:

- If $a \in Ap$, then a is a *PTLK* formula.
- If ϕ, φ are *PTLK* formulas, then $\neg\phi$, $\phi \wedge \varphi$ are *PTLK* formulas.
- If ϕ, φ are *PTLK* formulas, then $X^{\triangleright p}\phi$, $\phi \cup^{\triangleright p} \varphi$ are *PTLK* formulas, where $\triangleright \in \{>, \geq, <, \leq, =\}$, $p \in [0,1]$ is a real number.
- If ϕ is a *PTLK* formula, then $K_i^{\triangleright p}\phi$ with $i \in Ag$ is a *PTLK* formula.

In order to model random phenomena, we need to enrich transition systems with probabilities. We extend Kripke structure with probability for interpreting *PTLK*.

Definition 2. A probabilistic Kripke structure(*PK*) over a set of agents $Ag = \{1,\ldots, n\}$ and a set of propositions Ap is a 2n + 4 tuple $M = \{S, s_{0,} \sim_1,...,\sim_n, P, P_1,..., P_n, L\}$,where

- S is a finite set of states.
- $s_0 \in$ S is the initial state.
- $\sim_i \subseteq S \times S$ is an epistemic accessibility relation for each agent $1 \le i \le n$.
- $P : S \times S \to [0,1]$ is a transition probability function, such that for all s in S , $\sum_{s' \in S} P(s, s') = 1$.
- $P_i : S \times S \to [0,1]$ with $1 \le i \le n$ is a probability function such that $\sum_{s \sim_i s'} P_i(s, s') = 1$.
- $L : S \to 2^{Ap}$ is a labelling function assigning atomic propositions to states.

Before we give formal semantics to *PTLK*, we briefly review some concepts from probability theory. A nonempty set Ω of possible outcomes of an experiment of chances is called sample space. A set $B \subseteq 2^{\Omega}$ is called $\sigma-$ algebra over Ω if it contains Ω , $\Omega \setminus$ E for each E $\in$ B , and the union of any countable sequence of sets from B. The subsets of Ω that are elements of B are called measurable.

A probability space is a triple $PS = (\Omega, B, \Pr ob)$ where Ω is a sample space, B is an $\sigma-$ algebra over Ω , and $\Pr ob$ is a mapping B $\to$ [0,1] such that $\Pr ob(\Omega) = 1$ and $Prob(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \Pr ob(E_i)$ for any sequence $E_1, E_2, ...$ of pair disjoint sets from B . We call $\Pr ob$ a probability measure.

Definition 3(Paths). Let M be a *PK* .

- An infinite path π in M is an infinite sequence $s_0, s_1, \ldots$ of states such that $\forall i \ge 0$, $P(s_i, s_{i+1}) > 0$.
- A finite path π' is a finite prefix of an infinite path.

For a *PK* M , let $\Omega = S^{\omega}$ be the set of paths of M . Let B be the $\sigma-$ algebra generated by $\{C(\pi) \mid \pi \in S^*\}$, where $C(\pi) = \{ \pi' \in \Omega \mid \pi$ is a prefix of $\pi' \}$ is the basic cylinder set of π . A *PK* M induces a probability space. The underlying $\sigma-$ algebra is defined over the basic cylinder set induced by the finite paths starting in the initial state s_0 . The probability measure $\Pr ob$ is the unique measure on this $\sigma-$ algebra where $\Pr ob(C(s_0 s_1 ... s_n)) = \prod_{0 \le i \le n-1} P(s_i, s_{i+1})$.

We use $Paths_M^{\omega}(s)$ to represent the set of all infinite paths in M that start in state s . The subscript M is omitted when it is clear from the context. $Paths^{\omega}(s, \phi)$ is used

to represent the set of infinite paths that start in state s and satisfy ϕ. Formally $Paths^{\omega}(s,\phi) = \{\pi \in Paths^{\omega}(s) \mid \pi \models \phi\}$.Then $\Pr ob(s,\phi) = \Pr ob(\{\pi \mid \pi \in Paths^{\omega}(s,\phi)\})$. The semantics of *PTLK* are given via the satisfaction relation "|=" which holds between pairs of the form (M,s).

Definition 4(PTLK Semantics). Let M be a *PK* over the set *Ag* of agents and the set *Ap* of propositions, ϕ be a *PTLK* formula over *Ag* and *Ap*, and $s \in S$ be a state. Truth of ϕ at s in M written $(M,s) \models \phi$, is defined inductively as follows:

- For any atomic proposition $a \in Ap$, $M,s \models a$ if and only if $a \in L(s)$.
- $M,s \models \neg\phi$ if and only if $M, s \not\models \phi$.
- $M,s \models \phi \wedge \varphi$ if and only if $M,s \models \phi$ and $M,s \models \varphi$.
- $M,s \models X^{\rhd p}\phi$ if and only if $\sum_{M,s' \models \phi} P(s,s') \rhd p$, where $\rhd \in \{>,\geq,<,\leq,=\}$.
- $M,s \models \phi \bigcup^{\rhd p} \varphi$ if and only if $\Pr ob(s,\phi \bigcup \varphi) \rhd p$, where $\rhd \in \{>,\geq,<,\leq,=\}$.
- $M,s \models K_i^{\rhd p}\phi$ if and only if $\sum_{M,s' \models \phi \wedge s' \sim_i s} P_i(s,s') \rhd p$, where $\rhd \in \{>,\geq,<,\leq,=\}$.

3 Abstraction for Model Checking PTLK

3.1 Abstract Probabilistic Kripke Structure

Our aim is to provide an abstraction of *PK* which is conservative for some subsets of *PTLK* The basic principle is to collapse sets of concrete states into single abstract states such that the abstract system simulates the concrete system. The probability between abstract states is represented by an interval. The probability of epistemic accessibility relation is also represented with an interval. Taking the interval to represent the probability helps us define the under- and over-approximation.

Definition 5. An abstract probabilistic Kripke structure(APK) over a set of agents $Ag = \{1,...,n\}$ and a set of propositions *Ap* is a 3n+5 tuple $A = \{S, s_0, \sim_1, ..., \sim_n, P^l, P^u, P_1^l, P_1^u, ..., P_n^l, P_n^u, L\}$. where

- $S, s_0, \sim_1, ..., \sim_n, L$ are defined as in Definition 2.
- $P^l, P^u : S \times S \to [0,1]$ are matrices describing the lower and upper bounds for the transition probabilities between states such that for all $s,s' \in S$, $P^l(s,\bullet)$ and $P^u(s,\bullet)$ are pseudo distribution functions, where $P^l(s,\bullet) = \sum_{s' \in S} P^l(s,s')$, $P^u(s,\bullet) = \sum_{s' \in S} P^u(s,s')$, and $P^l(s,s') \leq P^u(s,s')$, and $P^l(s,S) \leq 1 \leq P^u(s,S)$.
- $P_i^l, P_i^u : S \times S \to [0,1]$ with $1 \leq i \leq n$ are matrices describing the lower and upper bounds for the probabilities over accessibility relation such that for all $s,s' \in S$, $P_i^l(s,\bullet)$ and $P_i^u(s,\bullet)$ are pseudo distribution functions,

where $P_i^l(s,\bullet) = \sum_{s' \sim_i s} P^l(s,s')$, $P_i^u(s,\bullet) = \sum_{s' \sim_i s} P^u(s,s')$, and $P_i^l(s,s') \le P_i^u(s,s')$, and $P_i^l(s,\bullet) \le 1 \le P_i^u(s,\bullet)$.

Let $H = (P, P_1, ..., P_n)$ be an n + 1 triple, where P is a probabilistic function between states , P_i with $1 \le i \le n$ are probabilistic functions over accessability relation. We call H is consistent with an APK $A = \{S, s_0, \sim_1, ..., \sim_n, P^l, P^u, P_1^l, P_1^u, ..., P_n^l, P_n^u, L\}$ if and only if for all $s, s' \in S$, $P^l(s,s') \le P(s,s') \le P^u(s,s')$, and for all $1 \le i \le n$, $s, s' \in S$ $P_i^l(s,s') \le P_i(s,s') \le P_i^u(s,s')$.

Definition 6 (Semantics for APK). Let $H = (P, P_1, ..., P_n)$ be an n + 1 triple of probabilistic functions, $A = \{S, s_0, \sim_1, ..., \sim_n, P^l, P^u, P_1^l, P_1^u, ..., P_n^l, P_n^u, L\}$ be an *APK*, and ϕ be a *PTLK* formula. If H is consistent with A , then we call $A, s_0 \models \phi$ over H, if and only if for the *PK* $M = \{S, s_0, \sim_1, ..., \sim_n, P, P_1, ..., P_n, L\}$, $M, s_0 \models \phi$.

Following the idea of abstraction in model checking reactive systems, we group states of the concrete system in the abstract system. Before defining the principle to induce an *APK* from a *PK*, we consider how to group the states. Intuitively, states of an abstract system should satisfy same propositions.

Definition 7. Let $M = \{S, s_0, \sim_1, ..., \sim_n, P, P_1, ..., P_n, L\}$ be a *PK*, $S' \subseteq S$, ϕ be a *PTLK* formula, A*tom*(ϕ) be a set of propositions occurring in ϕ . We call ϕ respects S' if for all $s, s' \in S'$, $L(s) \cap Atom(\phi) = L(s') \cap Atom(\phi)$.

Definition 8. Let $M = \{S, s_0, \sim_1, ..., \sim_n, P, P_1, ..., P_n, L\}$ be a *PK*, ϕ be a *PTLK* formula, $\Lambda = \{A_1, ..., A_m\} \subseteq 2^S$ be a partition of S , i.e., $A_i \ne \varnothing$, $A_i \cap A_j = \varnothing$ for $i \ne j, 1 \le i, j \le n$ and $\bigcup_{i=1}^{n} A_i = S$, and ϕ respect A_i for all $1 \le i \le m$. Then the abstraction of M induced by Λ and ϕ is the *APK* $A = \{\overline{S}, \overline{s_0}, \overline{\sim_1}, ..., \overline{\sim_n}, \overline{P}^l, \overline{P}^u, \overline{P_1}^l, \overline{P_1}^u, ..., \overline{P_n}^l, \overline{P_n}^u, \overline{L}\}$. given by

- $\overline{S} = \Lambda$.
- If $s_0 \in A_i$,then $\overline{s_0} = A_i$.
- For all $\overline{\sim_k}$ with $1 \le k \le n$, for all A_i, A_j with $1 \le i, j \le m$, $A_i \ \overline{\sim_k}\ A_j$ if and only if there exists $s \in A_i, s' \in A_j$ such that $s \sim_k s'$.
- For all A_i , A_j with $1 \le i, j \le m$, $\overline{P}^l(A_i, A_j) = \min_{s \in A_i} P(s, A_j)$, $\overline{P}^u(A_i, A_j) = \max_{s \in A_i} P(s, A_j)$, where $P(s, A_j) = \sum_{s' \in A_j} P(s, s')$.
- For all k with $1 \le k \le n$, for all A_i , A_j with $1 \le i, j \le m$, $\overline{P_k}^l(A_i, A_j) = \min_{s \in A_i} P_k(s, A_j)$, $\overline{P_k}^u(A_i, A_j) = \max_{s \in A_i} P_k(s, A_j)$, where $P_k(s, A_j) = \sum_{s' \in A_j} P_k(s, s')$.

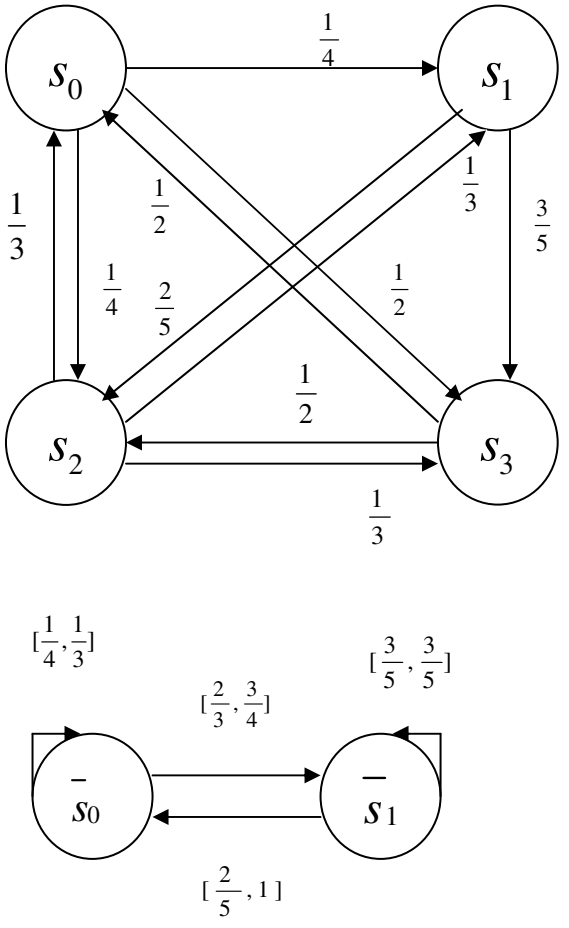


Fig. 1. Abstracting a Probabilistic Kripke Structure

- $\overline{L}: \overline{S} \to 2^{A_p}$ where $p \in L(A_i)$ if and only if there exists $s \in A_i$ such that $p \in L(s) \bigcap Atom(\phi)$.

Example 1. Let ϕ be a *PTLK* formula, and $Atom(\phi) = \{a, b\}$. Consider the *PK* in Figure 1.(left) with $Ap = \{a, b\}$, $Ag = \{1\}$, $L(s_0) = L(s_2) = \{a\}, L(s_1) = L(s_3) = \{b\}$, In Figure1. numbers labeled at the end of arrows represent the probability distribution between states. The probability function P_1 over accessibility relation is defined as follows: $P_1(s_0, s_1) = \frac{1}{3}$, $P_1(s_0, s_2) = \frac{2}{3}$, $P_1(s_2, s_1) = \frac{3}{4}$, $P_1(s_2, s_0) = \frac{1}{4}$, $P_1(s_1, s_0) = \frac{1}{2}$, $P_1(s_1, s_3) = \frac{1}{2}$, $P_1(s_3, s_2) = \frac{3}{5}$, $P_1(s_3, s_1) = \frac{2}{5}$. The *APK* induced by the partition:

$\{\{s_0, s_2\}, \{s_1, s_3\}\}$ and ϕ is depicted in Fig1.(right) with $\overline{L}(\overline{s_0}) = \{a\}$, $\overline{L}(\overline{s_1}) = \{b\}$. The probability functions $\overline{P}_1^l$ and $\overline{P}_1^u$ are defined as follows:

$$\overline{P}_1^l(\overline{s_0}, \overline{s_0}) = \frac{1}{4},\ \overline{P}_1^l(\overline{s_0}, \overline{s_1}) = \frac{1}{3},\ \overline{P}_1^l(\overline{s_1}, \overline{s_1}) = \frac{2}{5},\ \overline{P}_1^l(\overline{s_1}, \overline{s_0}) = \frac{1}{2},\ \overline{P}_1^u(\overline{s_0}, \overline{s_0}) = \frac{2}{3},$$

$$\overline{P}_1^u(\overline{s_0}, \overline{s_1}) = \frac{3}{4},\ \overline{P}_1^u(\overline{s_1}, \overline{s_1}) = \frac{1}{2}, \overline{P}_1^u(\overline{s_1}, \overline{s_0}) = \frac{3}{5},$$

3.2 Property Preservation Theorem

The aim of abstraction is to simplify the concrete system while preserving some properties. In this subsection we discuss which properties expressed with *PTLK* will be preserved.

Definition 9. Let $\triangleright \in \{<, \leq, =, >, \geq\}$. $PTLK_{\triangleright}$ is a subset of *PTLK* in which negation is applied only to atomic propositions. Formally, $PTLK_{\triangleright}$. Over $\triangleright$,a set Ag of agents, and a set Ap of propositions is defined by the following BNF expression: $\phi := a \mid \neg a \mid \phi \wedge \phi \mid \mathrm{X}^{\triangleright p}\phi \mid \phi \bigcup^{\triangleright p} \phi \mid K_i^{\triangleright p}\phi$.

Theorem 1. Let M be a PK, Λ be a partition of S , ϕ be a $PTLK_{\geq}$ ($PTLK_{>}$) formula, A be the abstraction of M induced by Λ and ϕ If $A, \overline{s}_0 \models \phi$ over $(\overline{P}^l, \overline{P}_1^l, ..., \overline{P}_n^l)$, then $M, s_0 \models \phi$.

Proof. The proof is by induction on the length of ϕ .

- Let $\phi = a$. Then by Definition 6, $A, \overline{s}_0 \models a$ if and only if $a \in \overline{L}(\overline{s_0})$. By the definition

of $\overline{L}$, $a \in \overline{L}(\overline{s_0})$ implies there exists $s \in \overline{s}_0$ such that $a \in L(s) \bigcap Atom(\phi)$. From Definition 7, $L(s) \bigcap Atom(\phi) = L(s_0) \bigcap Atom(\phi)$. So $a \in L(s_0)$, i.e. $M, s_0 \models a$.

- Let $\phi = \neg a$. Then by Definition 6, $A, \overline{s}_0 \not\models a$ if and only if $a \notin \overline{L}(\overline{s_0})$. By the definition

of $\overline{L}$, $a \notin \overline{L}(\overline{s_0})$ implies there are no state $s \in \overline{s}_0$ such that $a \in L(s) \bigcap Atom(\phi)$ Since $a \in Atom(\phi)$, $a \notin L(s_0)$, i.e. $M, s_0 \models \neg a$.

- Let $\phi = \varphi_1 \wedge \varphi_2$. By Definition 6, $A, \overline{s}_0 \models \varphi_1 \wedge \varphi_2$ implies $A, \overline{s}_0 \models \varphi_1$ and $A, \overline{s}_0 \models \varphi_2$. By

induction, $A, \overline{s}_0 \models \varphi_1$ implies $M, s_0 \models \phi_1$, $A, \overline{s}_0 \models \varphi_2$ implies $M, s_0 \models \phi_2$. That is $M, s_0 \models \varphi_1 \wedge \varphi_2$.

- Let $\phi = X^{\geq p}\varphi$. By induction, if $A, \overline{s}_0 \models \varphi$ then for all $s \in \overline{s}_0$, $M, s \models \varphi$. By the

definition of $\overline{P}^l$ we have $\sum_{s_1 \models \varphi} P(s_0, s_1) \geq \sum_{\bar{s}_1 \models \varphi} P(s_0, \bar{s}_1) \geq \sum_{\bar{s}_1 \models \varphi} \overline{P}^l(\bar{s}_0, \bar{s}_1)$. According to the semantics of $A, \bar{s}_0 \models X^{\geq p}\varphi$, $\sum_{\bar{s}_1 \models \varphi} \overline{P}^l(\bar{s}_0, \bar{s}_1) \geq p$, So, $\sum_{s_1 \models \phi} P(s_0, s_1) \geq p$. That is $M, s_0 \models X^{\geq p}\varphi$.

- Let $\phi = \varphi_1 \bigcup^{\geq p} \varphi_2$. Assume there exists a path $\bar{s}_0, ..., \bar{s}_n, ...$ in A such that $A, \bar{s}_n \models \varphi_2$, $A, \bar{s}_i \models \varphi_1$ for all $1 \leq i \leq n-1$. Direct from the definition of Prob and $\overline{P}^l$ we have $\Pr ob(\{s_0 ... s_n \mid 1 \leq i \leq n \wedge s_i \in \bar{s}_i\}) = \sum_{s_0 \in \bar{s}_0, ..., s_n \in \bar{s}_n} P(s_0, s_1) P(s_1, s_2) ... P(s_{n-1}, s_n)$ $= \sum_{s_0 \in \bar{s}_0, s_1 \in \bar{s}_1} P(s_0, s_1) \sum_{s_1 \in \bar{s}_1, s_2 \in \bar{s}_2} P(s_1, s_2) \ \ldots \ \sum_{s_{n-1} \in \bar{s}_{n-1}, s_n \in \bar{s}_n} P(s_{n-1}, s_n) \geq P(s_0, \bar{s}_1) ... P(s_{n-1}, \bar{s}_n) \geq \overline{P}^l(\bar{s}_0, \bar{s}_1) \ \ldots \ \overline{P}^l(\bar{s}_{n-1}, \bar{s}_n) = Prob\ (C(\bar{s}_0 ... \bar{s}_n))$. Further by induction, for each state $s_n \in \bar{s}_n$, $M, s_n \models \varphi_2$, and for all states $\bar{s}_i$ with $1 \leq i \leq n-1$, if $s_i \in \bar{s}_i$, then $M, s_i \models \varphi_1$. That is for each set of $C(\bar{s}_0 ... \bar{s}_n)$, there exists a corresponding set $\{ s_0 ... s_n \mid 1 \leq i \leq n \wedge s_i \in \bar{s}_i \}$ such that $\Pr ob(\{s_0 ... s_n \mid 1 \leq i \leq n \wedge s_i \in \bar{s}_i\}) \geq Prob\,(C(\bar{s}_0 ... \bar{s}_n))$. Therefore, $Prob(s_0, \varphi_1 U \varphi_2) \geq Prob(\bar{s}_0, \varphi_1 U \varphi_2) \geq p$.

- Let $\phi = K_i^{\rhd p}\varphi$. $A, \bar{s}_0 \models K_i^{\rhd p}\varphi$ means $\sum_{\bar{s} \models \varphi} \overline{P}_1^l(\bar{s}_0, \bar{s}) \geq p.$,

and $\sum_{s \models \varphi \wedge s_0 \sim_i s} P_i(s_0, s) \geq \sum_{\bar{s} \models \varphi} P_i(s_0, \bar{s}) \geq \sum_{\bar{s} \models \varphi} \overline{P}_i^l(\bar{s}_0, \bar{s})$. So $\sum_{s \models \varphi \wedge s_0 \sim_i s} P_i(s_0, s) \geq p$. That is $M, s_0 \models K_i^{\rhd p}\varphi$.

The following theorem states other *PTLK* formulas preserved by abstraction. Its proof is similar to Theorem 1. Therefore we omit its proof.

Theorem 2. Let M be a *PK*, Λ be a partition of S, A be the abstraction of M induced by Λ, ϕ be a $PTLK_{\leq}$ ($PTLK_{<}$) formula. If $A, \bar{s}_0 \models \phi$ with over $(\overline{P}^u, \overline{P}_1^u, ..., \overline{P}_n^u)$, then $M, s_0 \models \phi$ ..

4 Model Checking for PTLK

In this section, we present a model checking algorithm, which is given a probabilistic Kripke structure M and a *PTLK* formula ϕ determines whether $M, s_0 \models \phi$ .The algorithm is based on the algorithm for model checking in CTL. The algorithm will operate by labeling each state s with the set label(s) of subformulas of ϕ which are true in *s*. Initially, label(s) is just $L(s)$. The algorithm then goes through a series of stages. During the *i*th stage, subformulas with $i-1$ nested *PTLK* operators are processed. When a subformula is processed, it is added to the labeling of each state in which it is true. Once the algorithm terminates, we will have that $M, s_0 \models \phi$ if and only if $\phi \in label(s)$.

Algorithm 1. Check $X^{\triangleright p}\varphi$

$T := \{s \mid \phi \in label(s)\}$

while $S \neq \emptyset$ do

choose $s \in S$

$S := S \setminus \{s\}$

if $\sum_{s' \in T} P(s,s') \triangleright p$ then

$label(s) := label(s) \cup \{X^{\triangleright p}\varphi\}$

end if

end while

Fig. 2. Procedure for labeling the states satisfying $X^{\triangleright p}\varphi$

To handle formulas of the form $X^{\triangleright p}\varphi$, for each state s in M we first compute $P(s,\{s \mid \phi \in label(s)\})$. If $P(s,\{s \mid \phi \in label(s)\}) \triangleright p$, then we add $X^{\triangleright p}\varphi$ into $label(s)$. In Figure 2, we give a procedure *Check* $X^{\triangleright p}\varphi$ that adds $X^{\triangleright p}\varphi$ to $label(s)$ for every s that satisfies $X^{\triangleright p}\varphi$ assuming that ϕ has already been processed correctly.

Algorithm 2. *Check* $\phi \bigcup^{\triangleright p} \varphi$

$T := S$

while $S \neq \emptyset$ do

choose $s \in S$

$S := S \setminus \{s\}$

$Prob(s, \phi U \varphi) :=$

if $\varphi \in label(s)$ then 1

else if $\phi \notin label(s)$ then 0

else $\sum_{s' \in S} P(s,s') Prob(s', \phi \bigcup \varphi)$

end if

if $Prob(s, \phi U \varphi) \triangleright p$ then $label(s) = label(s) \cup \{\phi U^{\triangleright p}\varphi\}$

end if

end while

Fig. 3. Procedure for labeling the states satisfying $\phi \bigcup^{\triangleright p} \varphi$

For formulas of the form $\phi \bigcup^{\rhd p} \varphi$, Hans Hansson and Bengt Jonsson[10] have shown that $Prob(s_0, \phi U \varphi)$ satisfies the following recurrence equation:

$$Prob(s_0, \phi U \varphi) := \text{ if } \varphi \in label(s) \text{ then } 1$$
$$\text{else if } \phi \notin label(s) \text{ then } 0$$
$$\text{else } \sum_{s' \in S} P(s,s') Prob(s', \phi U \varphi)$$

The recurrence equation gives an algorithm shown in Figure 3. to label the state s with $\phi \bigcup^{\rhd p} \varphi$ if $Prob(s, \phi \bigcup \varphi) \rhd p$. For formulas of the form $K_i^{\rhd p} \varphi$, for each state s in M we first compute the set of states labeling with ϕ (denoted by T_ϕ) and the set of states $\sim_i$ which is equivalent to s(denoted by T_s).

Algorithm 3. Check $K_i^{\rhd p} \varphi$

$T_\phi := \{s \mid \phi \in label(s)\}$

while $S \neq \varnothing$ do

choose $s \in S$

$T_s = \{s' \mid s' \sim_i s\}$ }

if $\sum_{s' \in T_s \cap T_\phi} P_i(s,s') \rhd p$ then $lable(s) = label(s) \cup \{K_i^{\rhd p} \varphi\}$

end if

$S := S \setminus \{s\}$

end while

Fig. 4. Procedure for labeling the states satisfying $K_i^{\rhd p} \varphi$

If. $P_i(s, T_\phi \bigcap T_s) \rhd p$, then we add $K_i^{\rhd p} \varphi$ into $label(s)$. In Figure 4, we give a procedure *Check* $K_i^{\rhd p} \varphi$ that adds $K_i^{\rhd p} \varphi$ to $label(s)$ for every s that satisfies $X^{\rhd p} \varphi$

5 A Case Study

We illustrate the abstraction technique presented above on the verification of a security protocol: Chaum's Dining Cryptographers Protocol[11], which provides a mechanism for anonymous broadcast.

5.1 Dining Cryptographers Protocol

The protocol of the dining cryptographers was introduced in [11]. The original wording from [11] is as follows:

"Three cryptographers are sitting down to dinner at their favourite three-star restaurant. Their waiter informs them that arrangements have been made with the maitre d'hotel for the bill to be paid anonymously. One of the cryptographers might be paying for the dinner, or it might have been NSA (U.S. National Security Agency). The three cryptographers respect each other's right to make an anonymous payment, but they wonder if NSA is paying. They resolve their uncertainty fairly by carrying out the following protocol:

Each cryptographer flips an unbiased coin behind his menu, between him and the ryptographer on his right, so that only the two of them can see the outcome. Each cryptographer then states aloud whether the two coins he can see (the one he flipped and the one his left-hand neighbour flipped) fell on the same side or on different sides. If one of the cryptographers is the payer, he states the opposite of what he sees. An odd number of differences uttered at the table indicates that a cryptographer is paying; an even number indicates that NSA is paying (assuming that the dinner was paid for only once). Yet if a cryptographer is paying, neither of the other two learns anything from the utterances about which cryptographer it is."

It can be shown that after running this protocols, all the cryptographers are able to determine whether it was the NSA or one of the cryptographers who paid for dinner. Notice that similar versions of the protocol can be defined for any number of cryptographers greater than three. To formalize this protocol in our framework, let us first consider the specification of the problem. Introducing an integer variable pay to express that which cryptographer has paid for the dinner. Assume there are n cryptographers, the following formula expresses Chaum's requirements in the case of cryptographer 1:

$$\phi = (pay \neq 1 \rightarrow (K_1^{\geq \frac{2}{3}}(\bigwedge_{i=1}^{3} pay \neq i) \vee (K_1^{\geq \frac{2}{3}}(\bigvee_{i=2}^{3} pay = i) \wedge \bigwedge_{i=2}^{3} \neg K_1^{\geq \frac{2}{3}}(pay = i)))) .$$

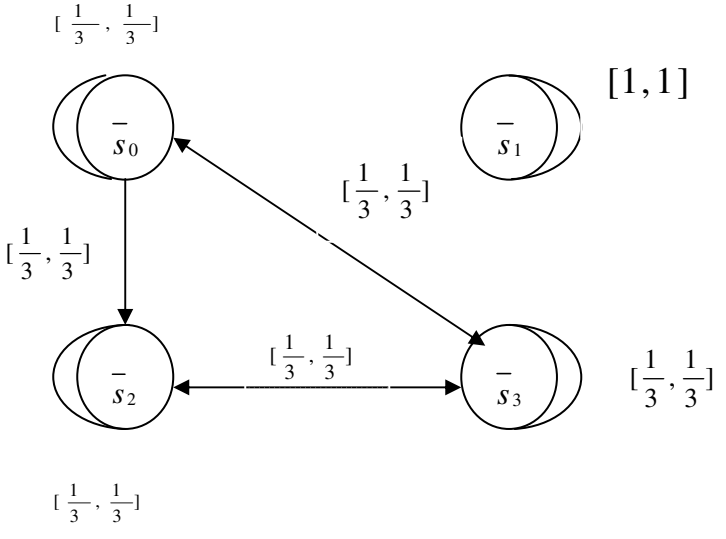


Fig. 5. Abstraction for protocol with three cryptographers

That is, if cryptographer 1 did not pay, then the he knows either the probability of that no cryptographer paid is no less than $\frac{2}{3}$, or knows the probability of that one of the other 2 paid, but does not know which is no less than $\frac{2}{3}$. The specifications for the other cryptographers are similar.

5.2 Abstraction

The abstraction model for the protocol with 3 cryptographers in shown in Figure 5. Since there are only four abstract states, it is easy to claim that the abstraction model satisfies the formula ϕ. That is the concrete model satisfies ϕ. In the concrete model there are 76 states, while the abstraction only has 4 states. Therefore, we have that the abstraction can reduce the model very much.

6 Conclusions and Future Work

We have presented a new probabilistic temporal logic of knowledge *PTLK* which not only can be used to reason about the uncertainty of knowledge, but also about probabilistic aspects of system's behaviors. In order to make model checking *PTLK* tractable, we proposed an abstraction technique to simplify concrete systems. There are many interesting avenues for future research. Our current work concentrates on providing evidence for the refutation of a *PTLK* property which is an essential feature of model checking, and using the refutation to guide the abstraction.

References

1. Clarke, E.M., Grumberg, O., Peled, D.: Model checking. MIT Press, Cambridge (2000)
2. Clarke, E.M., Grumberg, O., Long, D.E.: Model checking and abstraction. ACM Transactions on Programming Languages and Systems 16(5), 1512–1542 (1994)
3. Enea, C., Dima, C.: Abstractions of Multi-agent Systems. In: Burkhard, H.-D., Lindemann, G., Verbrugge, R., Varga, L.Z. (eds.) CEEMAS 2007. LNCS (LNAI), vol. 4696, pp. 11–21. Springer, Heidelberg (2007)
4. Cohen, M., Dam, M., Lomuscio, A., Russo, F.: Abstraction in model checking multi-agent systems. In: Proc. of 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 945–952 (2009)
5. de Ferreira, N.C., Fisher, M., van der Hoek, W.: Practical Reasoning for Uncertain Agents. In: Alferes, J.J., Leite, J. (eds.) JELIA 2004. LNCS (LNAI), vol. 3229, pp. 82–94. Springer, Heidelberg (2004)
6. Cao, Z.: Model Checking for Epistemic and Temporal Properties of Uncertain Agents. In: Shi, Z.-Z., Sadananda, R. (eds.) PRIMA 2006. LNCS (LNAI), vol. 4088, pp. 46–58. Springer, Heidelberg (2006)
7. Penczek, W., Lomuscio, A.: Verifying epistemic properties of multi-Agent systems via bounded model checking. Fundamenta Informaticae 55(2), 167–185 (2003)

8. Penczek, W., Lomuscio, A.: Verifying epistemic properties of multi-agent systems via bounded model checking. Fundamenta Informaticae 55(2), 167–185 (2002)
9. Su, K., Sattar, A., Luo, X.: Model Checking Temporal Logics of Knowledge Via OBDDs. Comput. J. 50(4), 403–420 (2007)
10. Hansson, H., Jonsson, B.: A Logic for Reasoning about Time and Reliability. Formal Aspects of Computing 6, 512–535 (1994)
11. Chaum, D.: The dining cryptographers problem: unconditional sender and recipient untraceability. J. Cryptology 1(1), 65–75 (1988)

Research on Double-Objective Optimal Scheduling Algorithm for Dual Resource Constrained Job Shop

Li Jingyao[1] Sun Shudong[1] Huang Yuan[2], and Niu Ganggang[1]

[1] Key Lab of Contemporary Design and Integrated Manufacturing Technology, MOE, 710072 Xi'an, China
[2] School of Mechatronics, Northwestern Polytechnical University, 710072 Xi'an, China
ljyao.6106@163.com, sdsun@nwpu.edu.cn, huangy@nwpu.edu.cn, ganggang.niu@hotmail.com

Abstract. To solve the double-objective optimization of dual resource constrained job shop scheduling, an inherited genetic algorithm is proposed. In the algorithm, evolutionary experience of parent population is inherited by the means of branch population supplement based on pheromones to accelerate the convergence rate. Meanwhile, the activable decoding algorithm based on comparison among time windows, the resource crossover operator and resource mutation operator, which are all established based on four-dimensional coding method are utilized with reference to the character of dual resource constrained to improve the overall searching ability. Furthermore, the championship selection strategy based on Pareto index weakens the impact of the Pareto level of chromosomes obviously. The elitist preservation strategy guarantees reliable convergence of the algorithm. Simulation results show that the performance of the proposed inherited GA is effective and efficient.

Keywords: dual resource constrained, inherited GA, time windows comparison, Pareto index.

1 Introduction

The Job Shop Scheduling Problem (JSP) is one of the most important issues in current academia, however, most of the literature on JSP had considered only machine as a limiting resource and ignored the possible constraints imposed by the availability of workers with requisite skills to perform the operations. This type of problem, where both machines and workers represent potential capacity constraints, was referred to as Dual-Resource Constrained (DRC) JSP by Nelson[1] at 1967. Since then there had been considerable investigations of DRCJSP which can be clustered into two groups and are briefly surveyed below. The first group investigated influences of different worker assignment rules such as "when" rule, "where" rule, "Push/Pull" rule using simulation method which forms the most investigated aspect of DRCJSP [2-4] [5-6]. The other group of studies investigated a great variety of heuristic algorithms and intelligent algorithms in order to find optimal or near optimal solutions to DRCJSP problems (e.g., Genetic Algorithm(GA) [7-10] and Ant Colony Optimization(ACO) [11]). However, most of the existing studies are single-objective optimization.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 222–229, 2010.

Ant Colony Optimization (ACO) is a constructed meta-heuristic algorithm based on swarm intelligence, in which artificial ants are created to solve problems. Since the initial work of Dorigo [12] on the ACO algorithm, several scholars have developed different ACO algorithms that performed properly when solving combinatorial problems especially in the field of JSP [13-14] in recent years. This paper presents the development of a hybrid algorithm which is the combination of ACO and Genetic algorithm (GA) for solving double-objective DRCJSP-HW.

2 Double-Objective Inherited Genetic Algorithm Introduction

In this paper, a double-objective inherited Genetic Algorithm (DOIGA) is proposed for solving scheduling problem in a dual-resource constrained environment. The algorithm uses four-dimension coding chromosomes, follows by the generation of branch population based on pheromones accumulation, and applies common genetic operators and resource evolutional operators to improve schedules. The Pareto-optimal solutions are selected based on taxis strategy with Pareto index.

2.1 Coding and Active Decoding

The representation of schedules as chromosome is important in the algorithm. In DOIGA, the model employs a list of operations for each part and associated resources including machines and workers as schedule representation, as shown in figure 1. Each gene G_i is formed by four elements G_i^p(p=1,2,3,4) and the length of the chromosome is equal to the total number of operations.

Part	…	1	1	3	2	…	1	2	…
Operation	…	2	3	3	4	…	4	5	…
Machine	…	1	5	2	3	…	2	3	…
Worker	…	3	3	2	3	…	3	4	…

Fig. 1. Four-Dimensional Coding Method

The fitness for a particular schedule is calculated with the following active decoding algorithm via comparison among time windows in view of the character of dual resource constrained environment. Some important definitions are put forward at first.

Definition 1. The intersection of time windows.

When time windows A [t_A^S, t_A^E] and B [t_B^S, t_B^E] satisfy $t_A^S < t_B^E \& t_A^E > t_B^S$, there is the intersection C [max(t_A^S, t_B^S),min(t_A^E, t_B^E)] which is recorded as $A \bigcap B = C$.

Definition 2. The inclusion of time windows.

B includes A, which is recorded as $B \supset A$, only if $A \bigcap B = A$.

Definition 3. The intersection of time window sets.

If there are two time window sets $C_1^T = \bigcup_{i=1}^{n_i} A_i$ and $C_2^T = \bigcup_{j=1}^{n_j} B_j$ without intersections among A_i or B_j, C_3^T is the intersection of C_1^T and C_2^T when $C_3^T = \{\bigcup_{i=1}^{n_i} C_i \mid C_i = \bigcup_{j=1}^{n_j} (A_i \cap B_j) = A_i \cap C_2^T\} \neq \varnothing$.

Based on these definitions, the steps of active decoding algorithm are following:

(1) Beginning with C_{P} which saves finish time of all parts' currently operations while C_{W} and C_{M} includes the available time window sets of all workers and machines, respectively.

(2) For each gene, read its process time $t^p_{G_i^2}$ and the finish time $t^e_{(G_i^2-1)}$ of its previous job ($t^e_{(G_i^2-1)}=0$ when $G_i^2 = 1$) from C_{P};

(3) Obtain the intersection C_R^T of associated machine G_i^3 and worker G_i^4 according to definition 3.

(4) The process time window of job G_i^2 is $[\max(t^S_A, t^e_{(G_i^2-1)}), \max(t^S_A, t^e_{(G_i^2-1)}) + t^p_{G_i^2}]$ when $A = \{\min t^S_A \mid \exists A[t^S_A, t^E_A] \in C_R^T \wedge A \supset [\max(t^S_A, t^e_{(G_i^2-1)}), \max(t^S_A, t^e_{(G_i^2-1)}) + t^p_{G_i^2}]\}$;

(5) Update C_{P}, C_{W} and C_{M} respectively and schedule the next operation until there is no unscheduled operations.

2.2 Crossover and Mutation Operators

The job scheduling evolution process utilizes composition operators to strengthen the diversity of search behavior. The crossover operators includes order crossover, linear order crossover and substring exchange crossover, and the mutation operators are swap mutation and insert mutation. On the other hand, the resource crossover operator comprises machine crossover and worker crossover which are executed with half population separately. The former is implemented by choosing two crossover points with the same machine at random, then exchanging the segments of process information (both G_i^1 and G_i^2) of the two genes, as shown in figure 2. The latter exchanges the segments of two genes with the same worker and needs to reset illegal resource which against resource ability constraint, as shown in figure 3. After that, the resource mutation operator which resets the resources allocation of random gene is implemented as shown in figure 4.

The evolution above may produce illegal chromosomes due to the strong coupling relationship among part, process and resources. Therefore a fast reconfigurable algorithm of illegal solution is proposed as following:

(1) Find $G_j = \min\{G_j^2 \mid G_j^1 = G_i^1 \wedge G_j^2 < G_i^2, j > i\}$;

(2) If $G_j \neq \varnothing$ for each gene G_i(i=1,2,...,n), exchange G_i and G_j, as shown in figure 5.

(3) If there is i<n, i=i+1 and go to step 1).

GA possesses the search capability which combines both breadth and depth as a result of the combination of crossover operator and mutation operator. However, the mutation probability is usually much less than the crossover probability which avoids egregious variation of elite but also limits the global search ability of the algorithm. Consequently, the fully evolution strategy where the probability of crossover and mutation are both equal to 1 while evolution with no weeded out is employed in this research to introduce genetic diversity into population and regain the information lost in earlier generations.

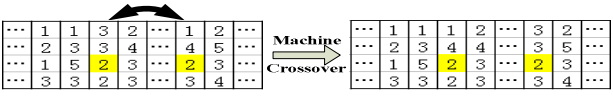


Fig. 2. Machine Crossover Operator

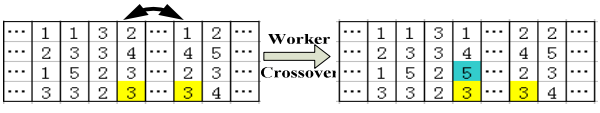


Fig. 3. Worker Crossover Operator

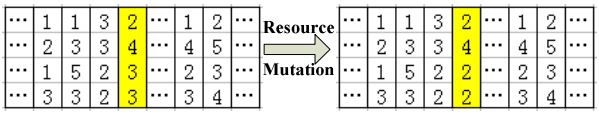


Fig. 4. Resource Mutation Operator

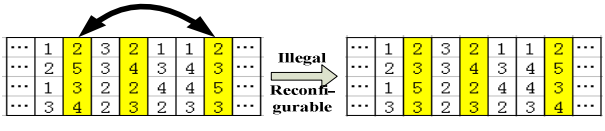


Fig. 5. Reconfigurable Algorithm

2.3 Sorting Strategy and Selection Operator on Pareto Index

Above all, the definitions of Pareto index and dominated relationship based on double-objective are proposed as following.

Definition 4. Chromosome A dominates B, which is recorded as $A \rhd B$, only if $(f^{O_1}(A) \leq f^{O_1}(B) \wedge f^{O_2}(A) < f^{O_2}(B)) \vee (f^{O_1}(A) < f^{O_1}(B) \wedge f^{O_2}(A) \leq f^{O_2}(B))$.

Definition 5. Pareto index is a three-dimension vector (N_g^A, N_e^A, N_b^A) which is calculated with $\{N_g^A = |C_j|, N_b^A = |B_i|, N_e^A = Num_c - N_g^A - N_b^A \mid \{B_i\} \lhd A \lhd \{C_j\}\}$ and employed to describe the domination relationship of A with the others, in which Num_c dedicates the total number of chromosomes participating in comparison. At first, the overlapped chromosomes in objective space should be reserved only one to avoid the premature problem caused by rapidly breed of elitist individual after active decoding process. Then each chromosome obtains its Pareto index according to comparison with the others which needs time complexity of O (N2+N). The whole population is divided into evolutionary set $P_{evolution}$ and Pareto-optimal set P_{Pareto} which is all the chromosomes with $N_g = 0$. At last the nearly optimal population with scale of N is selected from the $P_{evolution}$ based on the championship selection operator with Pareto index and elitist preserve strategy:

(1) Choose two chromosomes randomly;
(2) Compare their N_g, N_e, N_b in succession and select the smaller one;
(3) If the number of selected chromosomes is less than N, go back to step 1);
(4) Replace ten chromosomes of the selected population with larger N_g using elitist of $P_{evolution}$.

2.4 Knowledge Inherited Based on Pheromone

In order to improve the convergence performance of GA, DOIGA considers each technological process as branch path and introduces the pheromone as the carrier of knowledge inheriting to reflect the influence of them on different index. Before each iterative a branch population is generated based on pseudo-random proportional state transition rule with the pheromone concentration and some heuristic rules to inherit the evolution experience of parent population.

In general, ant colony select path according to both pheromone concentration and heuristic information, however, ant flow is another important but always be neglected impact factor in accordance with the ant moving experiment on symmetrical bridge of different branch wide, as shown in figure 6[15].

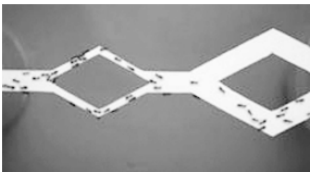


Fig. 6. Experiment of Dussutour

Consequently, the improved pseudo-random proportional state transition rule includes the information of ant flow, as shown in (1). The P_{P_x} represents the probability of being chosen of technological process P_x; the transfer probability q is

the important factor on balancing search-utilize performance of algorithm; $\tau_{P_x}^{O_1}$, $\tau_{P_x}^{O_2}$ means pheromone of P_x for time index and cost index respectively; η_{P_x} is the heuristic information generated with heuristic rules while n_{P_x} is the ant flow.

$$P_{P_x}=\begin{cases}1 & P\le q \wedge \arg\max_{P_x\in S_0}\{(\tau_{P_x}^{O1}\times\tau_{P_x}^{O2})^{\alpha}\times\eta_{P_x}^{\beta}\times n_{P_x}^{\gamma}\} \\ \dfrac{(\tau_{P_x}^{O1}\times\tau_{P_x}^{O2})^{\alpha}\times\eta_{P_x}^{\beta}\times n_{P_x}^{\gamma}}{\sum_{P_x\in S_0}((\tau_{P_x}^{O1}\times\tau_{P_x}^{O2})^{\alpha}\times\eta_{P_x}^{\beta}\times n_{P_x}^{\gamma})} & P>q\end{cases} \tag{1}$$

Depending on the evolution result $P_{evolution}$ and P_{pareto}, both local update and global update strategy of pheromone are utilized to update the pheromone and ant flow of technological process sets. First, the reasonable pheromone delta factor Q^{O_i} for different index is obtained with formula (2) to avoid too fast pheromone accumulation caused by different dimensions of different index.

$$Q^{O_i}=10^{\left\lceil \log_{10}(f^{O_i}(R_C^1))\right\rceil-2} \tag{2}$$

- Local Update: Update the pheromone and flow of all technological process in chromosomes of $P_{evolution}$ and P_{pareto} as (3).

$$\begin{cases}\tau_{P_x}^{O_i}=(1-\rho)\tau_{P_x}^{O_i}+Q^{O_i}/f^{O}(R_C^w) \\ n_{P_x}=n_{P_x}+1\end{cases} \qquad P_x\in R_C^w, R_C^w\in P_{pareto}\cup P_{evolution} \tag{3}$$

- Global Update: Update the technological process in P_{pareto} as (4) to accelerate the convergence rate.

$$\tau_{P_x}^{O_i}=\tau_{P_x}^{O_i}+Q^{O_i}/f^{O}(R_C^w) \qquad P_x\in R_C^w \tag{4}$$

3 Simulation Experiment and Analysis

3.1 Evaluate Criteria on Pareto-Optimal Solution

Because of the nonequilibrium character of the elements in Pareto-optimal solution, the lack of acknowledged evaluate criteria of Pareto-optimal solution becomes a realistic problem. In order to compare the calculate performance of DOIGA with the others, an evaluate criteria based on Pareto index is proposed. First of all, the dominate relationship comparison among several Pareto-optimal solutions is executed as following:

(1) Initial the Pareto index of all elements in Pareto-optimal solutions;

(2) The elements of set $P_{pareto}^{i}\{A_x^i\}$ compares with the ones in $P_{pareto}^{j}\{B_y^j\}$(j>i) in succession.

(3) Update the Pareto index of the elements participating in comparison respectively.

After that, the schedules with $N_g = 0$ are called as global Pareto-optimal solution with $R_{GPareto}$ records the number of them. Pareto-optimal solution with larger $R_{GPareto}$ and smaller $N_g^P = \sum_{A_x^i \in P_{pareto}^i} N_g^{A_x^i}$ approximates much more to the real Pareto-optimal front. On the other hand, the magnitude of N_g^P is much larger than $R_{GPareto}$ when the scale of Pareto-optimal solutions participating in comparison is large, so $C_{pareto} = \frac{R_{GPareto}}{N_g^P / 10^{\lfloor \log_{10}(\overline{N_g^P}) \rfloor}}$ is proposed as the evaluation criteria.

3.2 Experiment Results

This paper utilizes the Random Weight Sum Ant Colony Optimal (RWSACO) and NS-EIGA which based on EIGA [7] using fast Pareto sorting strategy of Deb[16] as the comparing algorithms. These algorithms are used to calculate the 10 groups of random benchmarks for 20 times respectively and the results are compared, as shown in Table I which proves DOIGA can achieve all global Pareto solutions of all numerical examples except Pro4, meanwhile the least $\overline{N_g^P}$ and largest $\overline{C_{pareto}}$ validate the validity and advanced of DOIGA on solving DRCJSP again.

Table 1. Algorithms Comparing

Example	DOIGA			RWSACO			NS-EIGA		
	$\overline{R_{GPareto}}$	$\overline{N_g^P}$	$\overline{C_{pareto}}$	$\overline{R_{GPareto}}$	$\overline{N_g^P}$	$\overline{C_{pareto}}$	$\overline{R_{GPareto}}$	$\overline{N_g^P}$	$\overline{C_{pareto}}$
Pro1	1.65	1137.15	14.37	0.00	2598.55	0.83	0.00	2335.50	0.62
Pro2	1.90	601.70	36.62	0.00	2856.95	0.59	0.00	3068.50	0.39
Pro3	2.45	1150.55	16.85	0.00	2530.80	0.59	0.00	2457.75	0.59
Pro4	2.05	707.00	18.32	0.00	4417.20	0.30	0.00	3076.90	0.41
Pro4	1.80	2851.55	2.60	0.75	806.35	6.66	0.20	1360.80	1.41
Pro5	1.75	431.00	32.70	0.00	2547.95	0.55	0.00	2157.50	0.57
Pro6	2.50	658.25	26.72	0.00	3476.80	0.35	0.00	3675.35	0.32
Pro7	0.95	665.45	94.48	0.00	1396.50	1.82	0.00	1272.30	1.36
Pro8	1.35	431.30	18.86	0.00	2460.25	0.58	0.00	2625.65	0.57
Pro9	1.45	480.00	83.78	0.00	2743.55	0.56	0.00	3003.25	0.46
Pro10	1.65	1137.15	14.37	0.00	2598.55	0.83	0.00	2335.50	0.62

4 Conclusion

In this paper, a Double-Objective Inherited Genetic Algorithm (DOIGA) is developed to solve the production scheduling problem of double-objective optimization in dual-resource constrained system. Some new technologies aimed at the character of dual resource constrained are proposed, such as the active decoding algorithm based on time windows comparison, resource crossover operator and resource mutation

operator, evolution experience inheriting based on pheromone, a fast sorting strategy based on Pareto index, and so on. Subsequently, the DOIGA is compared with RWSACO and NS-EIGA while the result proved that the proposed hybrid algorithm is very useful for optimizing the production performance under the realistic condition imposed by both machines and workers availability constraints.

References

[1] Nelson, R.T.: Labor and Machine Limited Production Systems. Management Science 13, 648–671 (1967)

[2] Treleven, M.: A Review of the Dual Resource Constrained System Research. IIE Transactions 21, 279–287 (1989)

[3] Kher, H.V.: Examination of worker assignment and dispatching rules for managing vital customer priorities in dual resource constrained job shop environments. Computers & Operations Research 27, 525–537 (2000)

[4] Berman, O., Larson, R.C.: A queueing control model for retail services having back room operations and cross-trained workers. Computers & Operations Research 31, 201–222 (2004)

[5] Treleven, M.: The Timing of Labor Transfers In Dual Resource-Constrained Systems: ”Push” Vs. ”Pull” Rules*. Decision Sciences 18, 73–88 (1987)

[6] Salum, L., Araz: Using the when/where rules in dual resource constrained systems for a hybrid push-pull control. International Journal of Production Research 47, 1661–1677 (2009)

[7] ElMaraghy, H., et al.: Scheduling of manufacturing systems under dual-resource constraints using genetic algorithms. Journal of Manufacturing Systems 19, 186–201 (2000)

[8] Zhijun, S., et al.: Intelligent optimization for job shop scheduling of dual-resources. Journal of Southeast University (Natural Science Edition) 35, 376–381 (2005)

[9] Shujuan, L., et al.: Mix optimization scheduling approach for multi-resource job-shop. Journal of System Engineering 22, 551–555 (2007)

[10] Binghai, Z., et al.: Scheduling Algorithm ofFlexible Production System Based on Dual Resource. Journal of South China University of Technolog 36, 45–49 (2008)

[11] Zhigang, L., et al.: Multi-resource constrained job-shop optimization scheduling based on ant colony algorithm. Journal of System Simulation 19, 216–220 (2007)

[12] Dorigo, M.: Optimization, learning and natural algorithms [in Italian], PhD thesis, Dipartimento di Elettronica, Politecnico di Milano, Milan (1992)

[13] Huang, K.-L., Liao, C.-J.: Ant colony optimization combined with taboo search for the job shop scheduling problem. Computers & Operations Research 35, 1030–1046 (2008)

[14] Huang, R.-H.: Multi-objective job-shop scheduling with lot-splitting production. International Journal of Production Economics 124, 206–213 (2010)

[15] Dussutour, A., et al.: Optimal traffic organization in ants under crowded conditions. Nature 428, 70–73 (2004)

[16] Deb, K., et al.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6, 182–197 (2002)

MDA Compatible Knowledge– Based IS Engineering Approach

Audrius Lopata[1,2] and Martas Ambraziunas[3]

[1] Kaunas University of Technology, Information Systems Department,
Studentu St. 50, Kaunas, Lithuania
[2] Vilnius University, Kaunas Faculty of Humanities,
Muitines St. 8, Kaunas, Lithuania
[3] Valdoware Inc.
Audrius.Lopata@ktu.lt, Martas.Ambraziunas@gmail.com

Abstract. Enhancement of MDA process with Knowledge Base Subsystem is aimed to reduce risk of project failures caused by inconsistent user requirements caused by insufficient problem domain knowledge. The enhancement of IS development environment with Enterprise Knowledge Base is discussed in this article. The major parts of Knowledge Base Subsystem are Enterprise Meta-Model, Enterprise Model and transformation algorithms.

Keywords: Enterprise Knowledge Based Information System Engineering, Model Driven Architecture, Enterprise Model, Enterprise Meta- Model.

1 Introduction

The majority of IT project failures (about 68% [2]) are caused by inconsistent user requirements and insufficient problem domain analysis. Although new methods of information systems engineering (ISE) are being researched and developed, they are empirical in nature: the project models repository of CASE system is composed on the basis of enterprise problem domain. The problem domain knowledge acquisition process relies heavily on the analyst and user; therefore it is not clear whether the knowledge of the problem domain is adequate.

The expert plays a pivotal role in the problem domain knowledge acquisition process, and few formalized methods of knowledge acquisition control are taken into consideration. Currently, despite of existing tools and CASE systems, user requirement analysis largely depends on the expertise of system analyst and the user. The knowledge stored in repository of CASE tool is not verified through formalized criteria, thus it is necessary to use advanced data capture techniques that ensure iterative knowledge acquisition process during which missing or incorrect data elements are obtained and fixed.

OMG provides Model Driven Architecture (MDA[4]) approach to information systems engineering where MDA focuses on functional requirements and system architecture not on technical details only. Model Driven Architecture allows long-term flexibility of implementation, integration, maintenance, testing and simulation. It

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 230–238, 2010.

means that enterprise modeling and user requirements engineering stages of information system engineering life cycle are not covered enough by MDA yet. There is lack of formalized problem domain knowledge management and user requirements acquisition techniques for composition and verification of MDA models. In order to solve this problem enhancement of MDA approach by the best practices [5], [6], [7] of Knowledge Base IS engineering (including Enterprise Knowledge repository) can be used. The proposed enhancement will intellectualize MDA models composition process by improving their consistency and decreasing the influence of the empirical information in composition process. Knowledge Base Subsystem will ensure MDA models verification against formal criteria defined by Control Theory [8]. It will reduce risk of project failures caused by inconsistent user requirements and insufficient problem domain knowledge verification.

2 Related Works

After an analysis of enterprise modeling standards including CEN EN 12204 [5], CEN EN 40003 (CIMOSA) [6], UEML [7] is performed the following usage shortcomings of mentioned standards to particular Information System's Engineering steps (user requirements acquisition, analysis and specification) are defined:

- The content of enterprise models, presently used in CASE systems, is not verified according to formalized criteria. Although enterprise models are created on the basis of various notations (such as data flow diagrams, work flow models etc.), their composition is not verified by CASE systems with respect to defined enterprise domain characteristics such as relations among particular processes, functions, actors, flows and etc.
- One of the reasons why the composition of enterprise models (created in this way) is incomplete for the improvement of CASE methods are insufficient features of both empirical Enterprise Models and user requirement specifications.

In order to solve these problems a particular Enterprise Meta-Model [9] has been developed in Kaunas University of Technology, Department of Information Systems. Its internal structure is based on Control Theory and best practices of the above mentioned enterprise modeling standards.

Recently, several authors [9], [10] proposed EMM based Use Case models generation methods in order to decrease empirical nature of Information Systems Engineering process already. These methods are able to cover static aspects of particular MDA steps (creation of Computation and Platform Independent Models), though mentioned EMM is not MDA compatible enough to fully support MDA process yet.

3 Knowledge Base MDA Principles

Most of MDA related techniques [1] are based on empirically collected problem domain knowledge thus negative influences verification of user requirements specification against actual customer needs. In some cases, user requirements do not correspond to formal business process definition criteria, which have a negative

impact on the next stage of information system engineering process. This problem can be solved by implementing Knowledge Base Subsystem (with EMM developed on basis of Control theory [8]) to particular MDA technique.

The Enterprise Knowledge Based Subsystem consists of three parts: the Enterprise Meta–Model (EMM), the Enterprise Model (EM) and model transformation algorithms. EM is a representation of the structure, activities, processes, information, resources, people, behavior, goals, and constraints of a business, government, or other enterprises [13]. EMM is formalized method based specification that regulates the formation order of the EM. Model generation algorithms handle transformation among MDA models (CIM, PIM) and EM.

3.1 Application of Knowledge Base Engineering Principles to MDA

According to survey [1], leading MDA based ISE methodologies need improvement in the following areas: requirements engineering, CIM construction, system models verification against problem domain processes. These problems can be solved by enhancing the MDA approach with Knowledge Base Subsystem.

EM construction requires formal CIM structure. Although the existing numerous techniques [3] describe CIM construction procedures, most of them are not formalized enough that influences negative impact on the EM constructs and composition. Usage of modified workflow diagrams [10] and formalized questionnaires (by demand free form diagrams can be used as well) solves such shortcomings and properly supports the suggested method. Conceptual MDA architecture enhanced with Knowledge Base Subsystem is depicted in Figure 1.

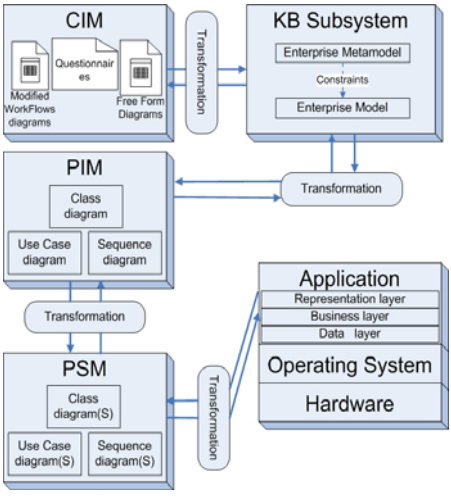


Fig. 1. Conceptual MDA architecture enhanced with Knowledge Base Subsystem

The set of modified workflow models [11] specifies the major CIM constructs. After CIM model is constructed, iterative verification process is started and is repeated until all the incorrect or missing CIM's elements are updated and corresponds to internal structure of EMM. The process leads to the creation of the

EM. The following step is the transformation of EM to PIM. PIM consist of the following UML diagrams: Class, Use Case and Sequence. According to IBM [15] these diagram types are essential to PIM and PSM. PSM is the data source for the Application code generation. The following layers of application internal structure can be generated from PSM: representation layer, business layer, data layer.

3.2 Mappings among the Internal Models of PIM and Enterprise Meta-model

During EM to PIM transformation process the following mapping types exists: EMM to Class Meta-Model (CMM), EMM to Use Case Meta-Model (UCM), EMM to Sequence Meta-Model (SMM). The class model of the Enterprise Meta-Model is defined in Figure 2.

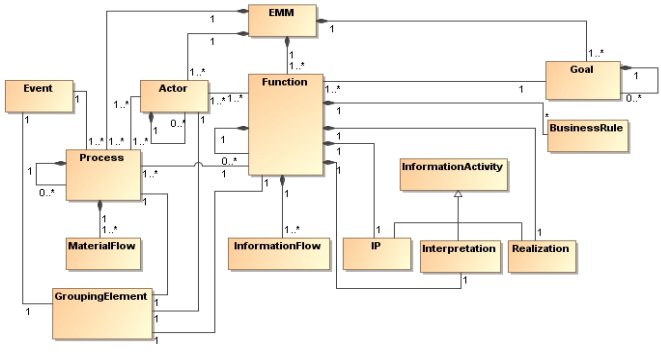


Fig. 2. Basic Elements of Enterprise Meta-Model

The class model of the Sequence Meta-Model is defined in Figure 3. Its core consists of nine classes, the basic of which are *Sequence*, *Message* and *Object*.

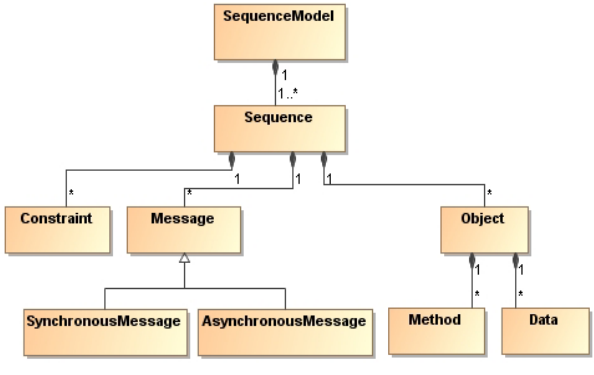


Fig. 3. Basic Elements of Sequence Meta-Model

Mapping the Enterprise Meta-Model onto the Sequence Meta-Model is depicted in table 1.

Table 1. Mapping among Enterprise Meta-Model and Sequence Meta-Model

SMM elements / EMM elements	Object	Data	Method	Messag.	Constr.
Function			Mapped	Mapped*	
Process			Mapped	Mapped*	
Material Flow		Mapped			
Information Flow		Mapped			
Actor	Mapped				
Event			Mapped	Mapped*	
Information Activity			Mapped	Mapped*	
Business Rule					Mapped

The class model of the Class Meta-Model is defined in Figure 4. Its core consists of eleven classes, the basic of which are *Class* and *Relationship*.

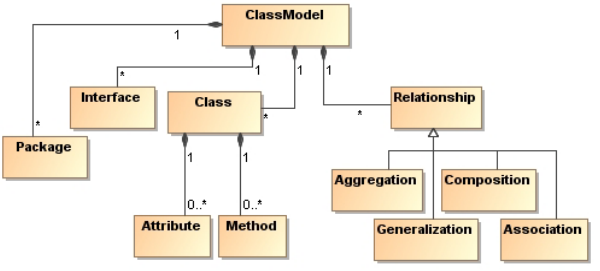


Fig. 4. Basic Elements of Class Meta-Model

Table 2. Mapping among Enterprise Meta-Model and Class Meta-Model

CMM elements / EMM elements	Class	Method	Attribut.	Relation	Interf.	Package
Function		Mapped				
Process		Mapped				
Material Flow			Mapped		Mapped	
Information Flow			Mapped		Mapped	
Actor	Mapped				Mapped	
Event		Mapped			Mapped	
Information Act.		Mapped				
BusinessRule				Mapped		
GroupingElement						Mapped

The class model of the Use Case Meta-Model is defined in Figure 5. Its core consists of six classes, the basic of which are *Actor*, *UseCase* and *ExtensionPoint*.

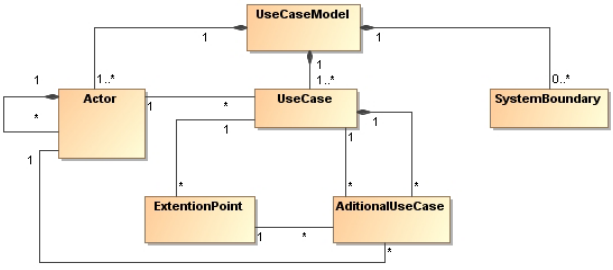


Fig. 5. Basic Elements of Use Case Meta-Model

Table 3. Mapping among Enterprise Meta-Model and Use Case Meta-Model

UMM elements / EMM elements	Actor	Use Case	Exten. Point	Additi UC	System Bound.	Pack.
Function		Mapped				
Process		Mapped				
Actor	Mapped				Mapped	
Event		Mapped				
Information Act.				Mapped		
Business Rule			Mapped			
GroupingElement					Mapped	Mapped

3.3 The Main Steps of Knowledge Base MDA Approach

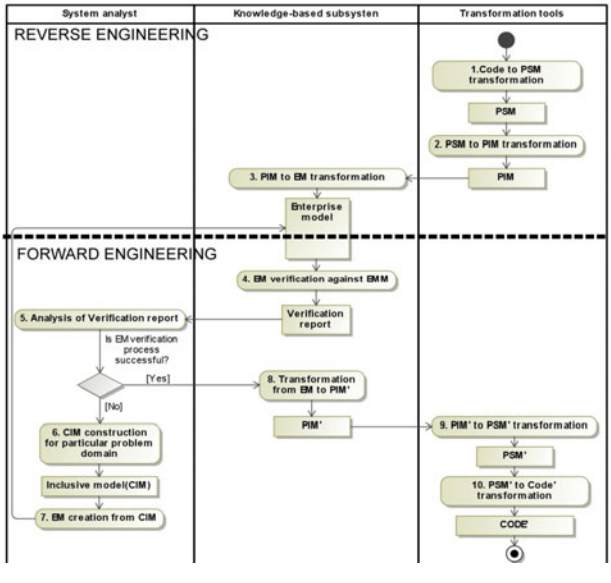


Fig. 6. Main steps of Knowledge Base MDA approach

Table 4. Detailed description of main steps of Knowledge Base MDA approach

STEP NAME	ACTOR	STEP DESCRIPTION	RESULT
1.Code to PSM transformation	Transformation tool	Particular MDA tool performs transformation from programming code (as well as to other artifacts such as executable files, direct link libraries) to PSM	**PSM**
2.PSM to PIM transformation	Transformation tool	Particular MDA tool performs PSM to PIM transformation process, removing or transforming platform related constructs to higher abstraction (PIM) level.	**PIM**
3.PIM to EM transformation	Knowledge Base Subsystem	Knowledge- Based subsystem transforms PIM (which basically consists of UML based models) to EM of particular problem domain.	**EM**
4.EM verification against EMM	Knowledge Base Subsystem	EM is verificated against Control Theory based EMM internal structure. Missing or incorrect EM data elements that do not correspond to EMM internal structure are determined during this step.	**Verification report**
5.Analysis of Verification report	System analyst	System analyst evaluates Verification report and approves transformation from EM to PIM' process in case of success or defines necessary actions in order to solve EM inconsistency against EMM internal structure issue in case of failure.	Identification of insufficient problem domain knowledge
6.CIM construction for particular problem domain	System analyst	Problem domain knowledge acquisition and CIM composition are performed at this step. This step is empirical in nature thus heavily depends on the system analyst's experience and qualification.	**CIM**
7.EM construction from CIM	System analyst	CIM using semi-automatic technique are transformed to EM.	**EM**
8.Transformation from EM to PIM	Knowledge Base Subsystem	XMI standard compatible PIM is constructed according to EM knowledge. It ensures PIM conformation to EMM defined formal constraints.	**PIM**
9.Transformation from PIM to PSM	Transformation tool	Particular MDA tool performs transformation from PIM to PSM, adding to PIM platform specific information.	**PSM**
10.Transformation from PSM to CODE	Transformation tool	Particular MDA tool performs transformation from PSM to programming code as well as to other artifacts such as executable files, direct link libraries, user documentation etc.	**CODE**

4 Conclusions

Knowledge Based Subsystem, which improves traditional MDA conception with best practices of problem domain knowledge and user requirements acquisition methods, is presented in this article. It ensures the problem domain knowledge verification against EMM internal structure. Such usage of Knowledge Based Subsystem together with MDA will improve the consistency of software artifacts and will reduce IT projects dependency on empirical processes.

The EMM is intended to be formal structure and set of business rules aimed to integrate the domain knowledge for the IS engineering needs. It is used as the "normalized" knowledge architecture to control the process of construction of an EM. Some work in this area has already been done [12], [14]. The EM is used as the main source of Enterprise Knowledge for discussed MDA approach.

References

1. Asadi, M., Ramsin, R.: MDA-based Methodologies: An Analytic Survey. In: Proceedings of the 4th European Conference on Model Driven Architecture, Berlin, pp. 419–431 (2008)
2. Ellis, K.: The Impact of Business Requirements on the Success of Technology Projects. Benchmark, IAG Consulting (2008)
3. Ambler, S.W.: Agile Modeling, http://www.agilemodeling.com/essays/inclusiveModels.htm
4. OMG: MDA Guide Version 1.0.1, http://www.omg.com
5. ENV 12 204. Advanced Manufacturing Technology Systems Architecture - Constructs for Enterprise Modelling. CEN TC 310/WG1 (1996)
6. ENV 40 003. Computer Integrated Manufacturing Systems Architecture - Framework for Enterprise Modelling, CEN/CENELEC (1990)
7. Vernadat, F.: UEML: Towards a Unified Enterprise modeling language. In: Proceedings of MOSIM 2001, Troyes, France, 2001-04-25/27, http://www.univ-troyes.fr/mosim01
8. Gupta, M.M., Sinha, N.K.: Intelligent Control Systems: Theory and Applications. The Institute of Electrical and Electronic Engineers Inc., New York (1996)
9. Gudas, S., Lopata, A., Skersys, T.: Approach to Enterprise Modeling for Information Systems Engineering. Informatica 16(2), 175–192 (2005)
10. Lopata, A., Gudas, S.: Enterprise model based computerized specification method of user functional requirements. In: International Conference 20th EURO Mini Conference EuroOpt-2008, Neringa, Lithuania, May 20-23, pp. 456–461 (2008) ISBN 978-9955-28-283-9
11. Lopata, A., Gudas, S.: Workflow-Based Acquisition and Specification of Functional Requirements. In: Proceedings of 15th International Conference on Information and Software Technologies IT 2009, Kaunas, Lithuania, April 23-24, pp. 417–426 (2009) ISSN 2029-0020
12. Gudas, S., Pakalnickas, E.: Enterprise Management view based Specification of Business Components. In: Proceedings of 15th International Conference on Information and Software Technologies, IT 2009, Kaunas, Technologija, pp. 417–426 (2009) ISSN 2029-0020

13. Leondes, C.T., Jackson, R.H.F.: Manufacturing and Automation Systems: Techniques and Technologies, p. 97. Academic Press, London (1992) ISBN 0120127458
14. Skersys, T.: Business Knowledge-Based Generation of the System Class Model. Informatica 37(2), 145–153 (2008)
15. Lloyd, A.: How to Migrate from Code-centric to Model-centric Development using Rational Software Archit, http://www.ibm.com/developerworks/rational/library/04/r-3247/index.html

ARIMA Signals Processing of Information Fusion on the Chrysanthemum

Lanzhou Wang[1] and Qiao Li[2]

[1] College of Life Sciences, China Jiliang University, Hangzhou 310018, Zhejiang, China
lzwang@cjlu.edu.cn
[2] College of Science, China Jiliang University, Hangzhou 310018, Zhejiang, China

Abstract. Weak electrical signals of the plant were tested by a touching test system of self-made double shields with platinum sensors. Tested data of electrical signals were denoised with the wavelet soft threshold. A novel autoregressive integrated moving average (ARIMA) model of weak electric signals of the chrysanthemum was constructed by the information fusion technology for the first time, that is, Xt =1.93731Xt-1 - 0.93731Xt-2 + εt + 0.19287εt–1 - 0.4173 Xt-2 - 0.17443 Xt-4 - 0.07764 Xt-5 - 0.06222 Xt-7. A fitting standard deviation was 1.814296. It has a well effect that the fitting variance and standard deviation of the model are the minimum. It is very importance that the plant electric signal with the data fusion is to understand self-adapting regulations on the growth relationship between the plant and environments. The forecast data can be used as preferences for the intelligent system based on the adaptive characters of plants.

Keywords: information fusion, ARIMA model, weak electrical signal, chrysanthemum.

1 Introduction

Now the information fusion is an effective technology to solve a lot of chaos information in various natural phenomena. The virtue of Autoregressive integrated moving average model (ARIMA model) is better ability to abstract the character and its model coefficient can make up of an eigenvector. We attempt to construct an ARIMA model of actual plant electric signals, and to obtain parameters of the model based on data fusion. The plant weak electric signal is the main physiology signal which is correlated to the plant physiological process of its inner transfer message for growth and it is a reaction to the stimulation of environments [1-9]. Electrical wave signals evoked the plant to come into being physiological changes, such as movement, metabolizing, growth and substance transportation etc., and to regulate relationships between the plant and the environment. If an appropriate model is founded, it can be applied to the research on self-adapting characters of plants, so as to develop some new instruments or methods for testing and determining their growth, metabolizing and substance transportation etc. and so as to realize the automatization management in crop/vegetable production to achieve the energy saving.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 239–247, 2010.

It is a basic method to construct a parameter model of signals for the signal processing [10-12]. Recently, it is quite common that the parameter model applied to the processing of the biomedical signal. One of the typical model is the ARIMA model in the parameter models, owing to easy calculation relatively for it, was tried to have application in the weak signal processing, such as brain wave signal [13-14], electromyographic signal and electrocardiosignal. Up to now, we cannot find a parameter model by the information fusion for processing plant signals. As to the no-stability signals [10-11] of time series in plant, some differences used to be a stabile series and then to express as a combination of the moving average both the autoregressive and noises on series. That is why the information fusion with ARIMA model is set up to analyze signals of the chrysanthemum.

2 Materials and Method

Aim at the difficult to process the plant no-stability signals of. An information fusion technology including signals test, denoise and process etc. will be used.

2.1 Testing Instruments and Plant

Chrysanthemum (*Dendranthema morifolium*) was purchased from the flower market and cultivated in lab surround. Testing instrument and method see reference 10-11 for detail.

2.2 Set Up a Testing Device and Analysis of Obtained Original Electric Signals of Plants

The basic principle of a testing device of plants: first, introducing the plant electrical signal by leading electrode into the biological testing system; second, magnifying and filtering electric signals; digitizing it by analog digital converting, and input the digitizing signal into the computer to analyze it (Fig. 1). Signals obtained in the experiment are de-noised by the wavelet soft-threshold de-noising method [10-11] in the MATLAB.

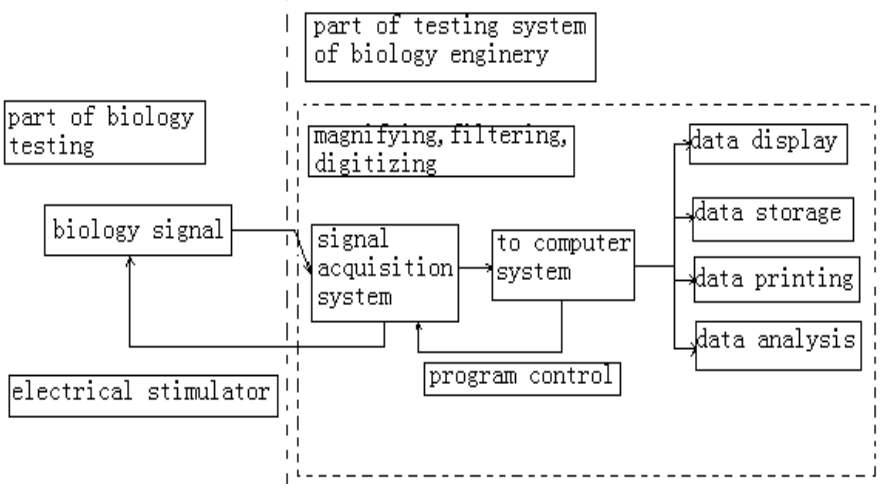


Fig. 1. Testing system of weak electrical signals in plants based on the information fusion

2.3 Time, Frequency Domain, Wavelet Transform Analysis

Basing on such characteristic, applying wavelet transform to the analysis of the plant electric wave signal to discuss the character of it. The wavelet transform comes from the analysis of signal, and has the ability of giving expression to time domain and frequency domain information at the same time. The function (basic wavelet), which satisfies some conditions, is used to decompose original signal. And the wavelet function make up of the wavelet basis by extending and contracting in frequency domain and translating in time domain. The definition of wavelet transform as follows:

$$WT_f(a,\tau) = < f(t), \Psi_{a,\tau}(t) > = \frac{1}{\sqrt{a}} \int_R f(t) \overline{\Psi\left(\frac{t-\tau}{a}\right)} dt$$

in the formula: a, denoting the scale frequency factor; t, denoting the time parameter; $\Psi(t)$, denoting the wavelet generating function. The wavelet transform $WT_f(a,\tau)$ is the two variable function, which was composed of the scale variable a and the variable τ. The variable a and variable τ defined the scale/ time plane. So, when we expand the function by the wavelet basis, it means that project a time function on the scale/ time plane. In this plane, we can sharply get the feature description of the signal, in different scale and time.

The most important application of the wavelet analysis is the denoising to the signal. The model of the noisy signal can be expressed as follow:

$$s(i) = f(i) + \sigma \cdot e(i) \quad i = 0, \cdots, n-1$$

$f(i)$ is the actual signal, $e(i)$ is the yawp, $s(i)$ is the noisy signal.

In the plant electric wave signal, the useful signal usually is low frequency signal, and the noisy signal usually is high frequency signal. So, the denoising course can be done by the following methods: firstly, decomposing the signal and the noise contain in the high frequency coefficient; secondly, processing the wavelet coefficient by the threshold value; then, reconstructing the signal to denoise. In this study, we choose the 'db3' wavelet as the basis function to decompose the plant electric wave signal in 5 levels, and got the low and high frequency coefficient, and reconstructed the coefficient to denoise.

2.4 Autoregressive Integrated Moving Average Model

As to the no-stability signals of time series, some differences should be used, to make into a stabile series and then to express as a combination of the moving average both the autoregressive and noises on the series from now to the past point. Autoregressive integrated moving average is abbreviated into model of ARIMA (p, d, q), which consists of the models at the autoregressive (AR), moving average (MA) and autoregressive moving average (ARMA). It shows as follow:

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$$

$\nabla^d = (1-B)^d$ is difference d; x_t = original series; ε_t = noise series; B =delay arithmetic operators.

Arithmetic operators of the AR: $\Phi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p$

Arithmetic operators of the MA: $\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$

Steps:

(1) The stabile checkout of original series: If it was no-stability, should be translated into a stabile series by the operation of difference.

(2) Understanding of models: It holds the approximate direction by reading of ACF (Self correlation coefficient) and PACF (Partial self correlation coefficient) charts for ranking of the aim sequence and to provide several models so as to perfect it.

(3) Parameter estimation and model diagnosis: The parameter estimation is to that the model diagnosis is by the model parameter estimation and hypothesis testing in the recognised phase.

(4) Forecast.

2.5 Program on ARIMA for Matlab

```
goptions reset=global;
gunit=pct cback=white border htext=5
ftext='style' colors=(black);
data number;
do number=1 to 15000 by 1;
output;
end;
data series;
merge number value;
proc print data=series; run; proc arima data=series;
identify var=F1(1) minic p=(0:24) q=(0:24);
estimate p=m q=n noint;
forecast lead=10 id=number out=results;
proc gplot data=results;
plot F1*number=1 forecast*number=2/noframe vaxis=axis1 haxis=axis2;
symbol1 c=black i=join v=none l=1;
symbol2 c=black i=join v=none l=1;
axis1 label=(a=90 'Amplitude(μV)');
axis2 label=('Number'); run;
```

3 Results and Discussion

3.1 The Time Domain Waveform of Electric Wave Signal

Fig.2 is a time domain waveform of electrical signals in the chrysanthemum. Electrical signals of the plant are a weak signal, varied acutely with the different time (Fig. 2 s). In the plant weak electric wave signal, the noisy signal is high frequency signals (Fig. 2 d1-d5), and the useful signal is low frequency signal after denoised (Fig. 2 a3).

Through calculating and trying, the fifth level decomposing is the best. So, that is why we choose the ‘db3’ wavelet as the basis function, to decomposing plant electric

signals in 3 levels (Fig. 2). The wavelet analysis is of the ability to express local characteristic of the signal in both time and frequency domain, and can accurately detect the arisen position of action electric waves. So, the wavelet analysis can be applied to research the plant no-stability electric signal as well.

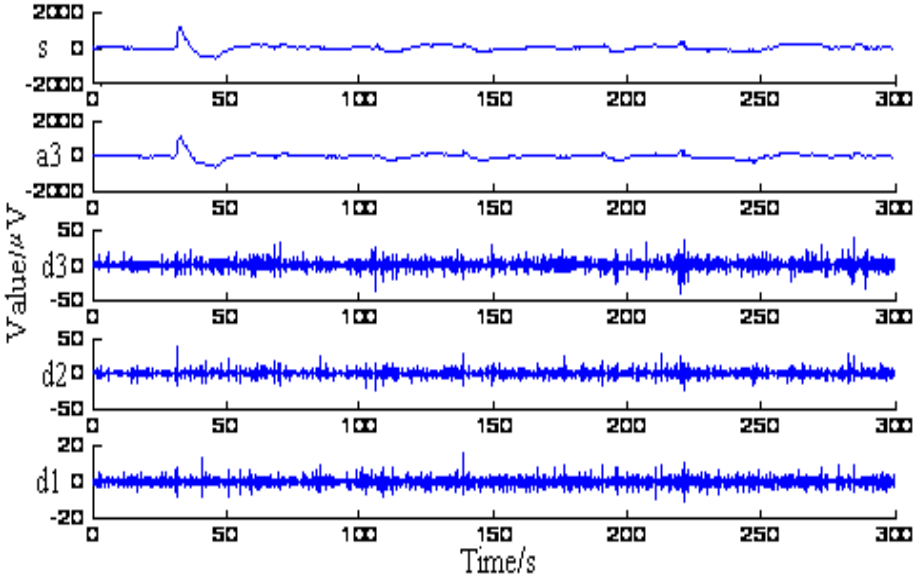


Fig. 2. Wavelet 3decomposing signals of the chrysanthemum (origin s, low frequency a3, high frequency d1-d3)

It shows that the signal in the chrysanthemum is a kind of complex weak electric signals including the action, variation and shock electric signals (Fig. 3: time 1-8).

An appearance of the action wave from the first to second time in the plant signals is about 30s and the action wave appears 8 times nearly in 300s, and that, action waves increasing as surge, the maximum of signals is about 1200 μv (Fig. 3).

3.2 The Information Fusion Forecast of Electric Signals

The chrysanthemum original electric wave signal (a piece of signal, same as follow) obtained in the experiment is Fig. 4a. The ARIMA model of the plant electric signal is shown that a fitting formula is X_t=1.93731X_{t-1}-0.93731 X_{t-2}+ε_t+ 0.19287ε_{t-1}-0.4173 X_{t-2}-0.17443 X_{t-4}-0.07764 X_{t-5}-0.06222 X_{t-7}. A fitting standard deviation is 1.814296. Fig. 4a and Fig. 4b are almost similarly from the model figure, which indicates that the ARIMA model is constructed successfully and can be used in forecast analysis for plant weak electric signals.

For testing the fitting effect of ARIMA model, we can forecast the tail 10 values of plant electric signals by using the founded model. The comparison between the true and forecast values is shown in Table 1. Results indicate that the effect of short period forecast is well.

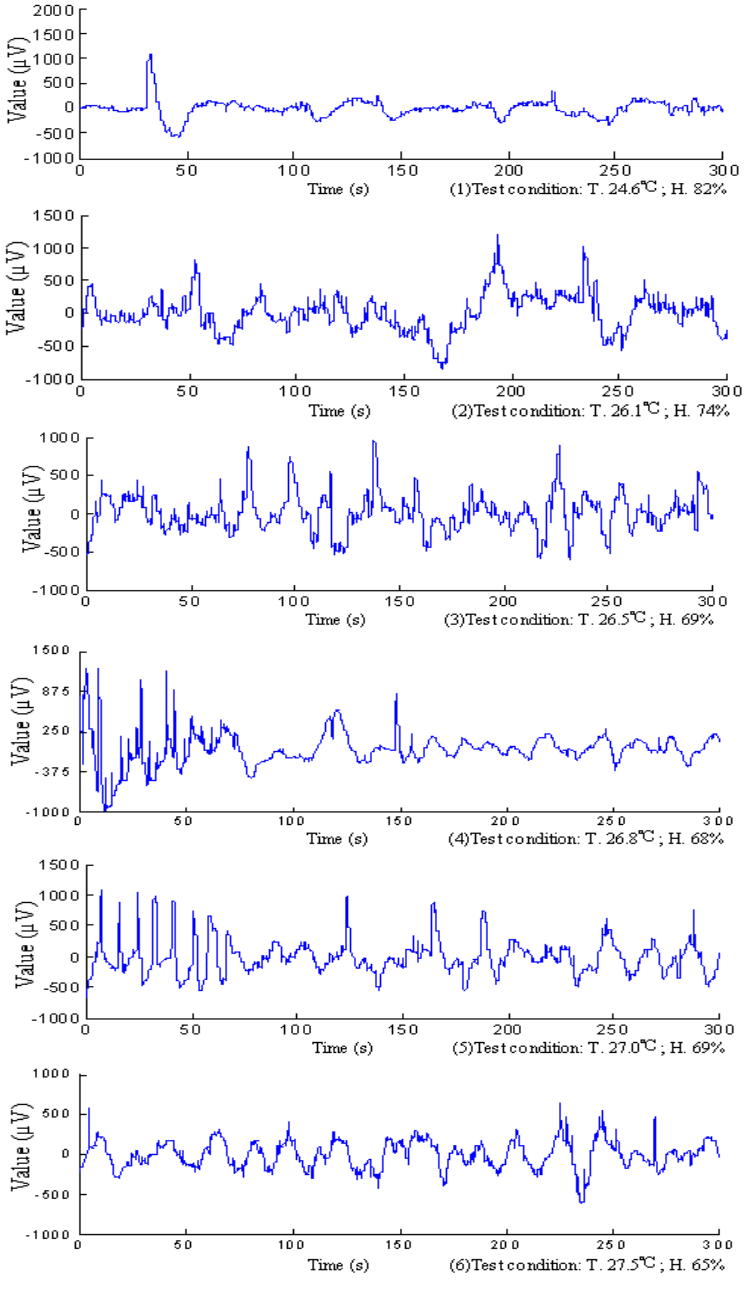


Fig. 3. Testing electrical signals after denoise in *D. pinnata* (each 5 min. in 8 times of a day at 0:00-8:00, 26 Mar. 2006).

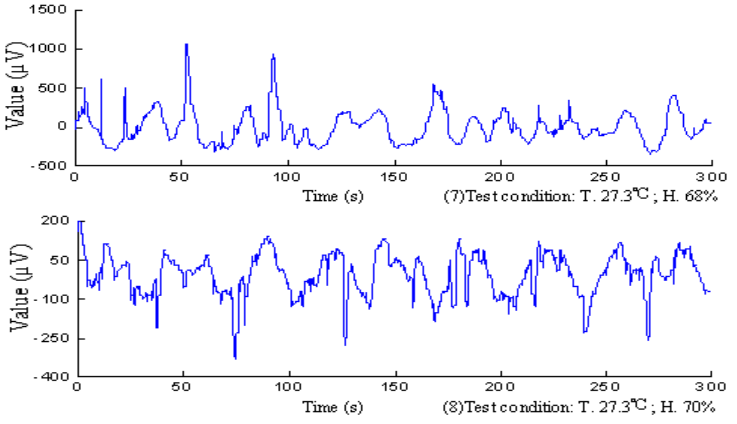


Fig. 3. (*continued*)

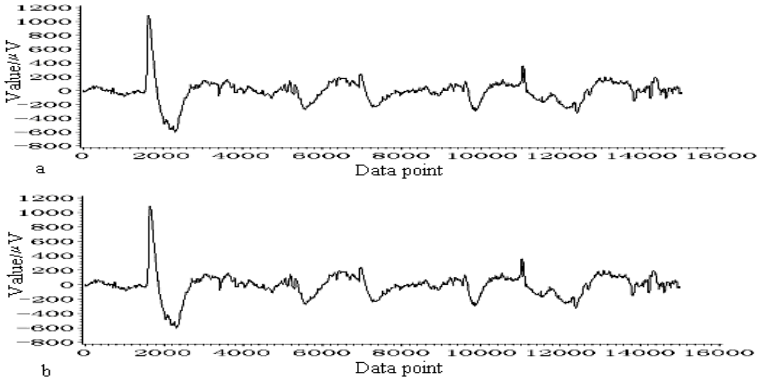


Fig. 4. Chrysanthemum original signals (a); fitting curve (b)

Table 1. Comparison between the true and predicted values of the signals of the ARIMA model in the chrysanthemum

Future points	1	2	3	4	5	6	7	8	9	10
True value	-23.551	-24.052	-24.988	-26.189	-25.777	-24.848	-24.267	-23.421	-23.348	-23.516
Forecast value	-23.306	-23.186	-23.063	-22.805	-22.540	-22.281	-22.023	-21.782	-21.556	-21.344
Error (%)	1.040	3.601	7.704	12.921	12.558	10.331	9.247	6.998	7.675	9.236

From the comparison results both the true value of the plant electric wave signal and the forecast values of ARIMA model in ten future positions, it shows that the forecast error of the beginning position is smaller and the forecast result is perfect, which indicates further that the plant ARIMA model we founded is successful. But with the

increasing of the future position, the forecast error based on the previous value is also increasing, which is one of limitation of the ARIMA model. Up to now, an application of data fusion at signals in the plant is finished.

4 Conclusions

It is not enough only understood characters of plant signals under a large numbers of chaos information. It is a good choice and succeeds in the information fusion technology for a research of plant weak electrical signals. There are abundant electrical signal information in the plant and it is difficulty to process and analyse the useful data from the disorderly and unsystematic information. So, the modern technology of weak signal processing should be used.

As a dynamic model for representing the plant electro-physiological signal, a construction of the ARIMA model by the information fusion makes the character of plant weak electric signals could be represented by the coefficients of the model. The fitting standard deviations are small, that is, the effect is well for a base of the application in future.

Based on the study we made before, obtained large numbers of data, it will construct a standard of ARIMA coefficient of plant signals and the environmental factor. We could compare the new monitoring data to forecast how the plant is growth in different conditions.

Acknowledgment. We thank that project subsidized by Zhejiang Povincial Natural Science Foundation of China under Grant No. Y3090142.

References

1. Mwesigwa, J., Collins, D.J., Volkov, A.G.: Electrochemical signaling in green plants: effects of 2, 4-dinitrophenol on variation and action potentials in soybean. Bioelectrochemistry 51, 201–205 (2000)
2. Alexander, G.V.: Green plants: electrochemical interfaces. Journal of Electroanalytical Chemistry 483, 150–156 (2000)
3. Gyenes, M.A., Kuerella, G.A.: Rhythmic excitation in Nitella at conditions of voltage. J. Exp. Bot. 34, 83–86 (1983)
4. Sakamoto, M., Sumiya, K.: The bioelectrical potentials of young woody plants. Wood Research 70, 42–46 (1984)
5. Paszewski, A., Zawadzki, T.: Action potentials in Lupinus angustifolius shoots 3 determination of the refractory periods. J. Exp. Bot. 27, 369–374 (1976)
6. Pickard, B.G.: Action potential in higher plants. Bot. Rev. 39, 172–201 (1973)
7. Paszewski, A., Zawadzki, T.: Action potentials in Lupinus angustifolius shoots. J. Exp. Bot. 24, 804–809 (1973)
8. Lou, C.H.: The messenger transmission of the electrochemistry wave in higher plant. Acta Bioph. Sinica 12, 739–745 (1996)
9. Li, H.X., Wang, L.Z., Li, Q.: Study on electric signals in Clivia miniata. Journal of China Jiliang University 16, 62–65 (2005)

10. Wang, L.Z., Li, H.X., Lin, M.: Analysis of plant electrical signal in the time domain and frequency domain. J. China Jiliang University 16, 294–298 (2005)
11. Lu, L.: The modeling and forecast of the non-stable time series state space. The Journal of The China Economic Research Center in Peking University 10, 1–9 (1999)
12. Ji, Z., Qin, S.R., Peng, L.L.: Signal processing of electroencephalogram and it's application. Journal of Chongqing University 25, 108–112 (2002)
13. Graupe, D.: Function separation of EMG signals via ARMA identification methods for prosthesis control purposes. IEEE Trans. Syst. Man Cybern. 5, 252–259 (1975)
14. Peter, J.B., Richard, A.D.: Time Series: Theory and Methods. M. Springer Publishing Group, Heidelberg (2001)

Noise Uncertainty Study of the Low SNR Energy Detector in Cognitive Radio

Guoqing Ji and Hongbo Zhu

Nanjing University of Posts and Telecommunications,
province key laboratory on wireless communicantion of Jiangsu, Nanjing, China
ji-guoqing@hotmail.com, Wilab@njupt.edu.cn

Abstract. It's well known that a signal-noise-ratio(SNR) threshold named "SNR wall"[5] appears in the energy detection due to the noise uncertainty. It makes energy detector(ED) highly non-robust in low SNR environment. In this paper, a log-normal distribution model of the noise power uncertainty is proposed. The detection performances of the energy detector are analyzed based on the proposed noise model. The numerical results illustrate that not only local but also cooperative low SNR Eds are reduced to invalid for a biggish noise power uncertainty.

Keywords: cognitive radio; energy detector; noise uncertainty; log-normal distribution.

1 Introduction

The USA Federal Communications Commission (FCC) has approved secondary usage of TV bands for mobile communications[2].Cognitive radio is considered as the promising technology for these and other shared spectrum applications[1].IEEE has already founded an 802.22 workgroup to establish the standards of WRAN based on CR techniques[3].Distinguishing the absence or presence of the primary user become principal tasks of the congnitive users. Spectrum sensing is the key technology in Cognitive Radio (CR) for primary user detection. Due to its generality and low complexity, energy detector(ED) becomes the most widely used detector and has been studied extensively[4,7,9]. However, spectrum sensing is always a challenging task practically. It's delightful to the secondary users(CU) when the SNR is sufficient. But mostly, the actual SNR for detection may be very low, and the received primary user's signals are deeply attenuated even when the CU (the detection node) is nearby or the CU is very far away inherently. However, the CU still needs to detect the primary user to avoid strongly interfering with the primary receiver unadvisedly if it transmits. Furthermore, the negative effects of the noise uncertainty on the detector performance becomes prominent due to the low SNR[5,6,7]. The uncertainty comes from the independent noise sources and the recever's approximation[5]. R. Tandra and A. Sahai[5,6] have made deeply studies on the noise uncertaity in low SNR "monent detector" on the basement of H. Urkowitz [4]and other antecessors,

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 248–255, 2010.

proposing the "SNR wall" conception. Yonghong Zeng [8] etc. proposed an uniformly noise power uncertainty distribution model. Wei Lin and Qinyu Zhang [9] proposed a discrete/ continuous model on the noise uncertainty, which we think it stopgap. We propose a log-normal distributed model on the noise uncertainty statistic characteristics, which is more legitimated based on the statistic theory. Based on the model, our studies show that cooperative sensing can alleviate the negative impacts of the noise uncertainty in some sort, but is not all-purpose here.

The remainder of this paper is organized as follows: section 2 first reviews Energy detector(ED) model and propose our noise power uncertainty model. In section 3, performances of single and cooperated detectors are formulated and numerical analyzed based on the proposed model. Conclusions and further works are provided in section 4.

2 System Model

2.1 Energy Detector

The formulations of the energy detector have had a set of forms since H.Urkowitz[4]. The path gain is often assumed unitarily nowadays. The amount of the sensors/receivers in network is assumed as $M \geq 1$. The sensing problem for each detector is formulated into the following binary hypothesis testing $\mathcal{H}_0 : x_i(n) = w_i(n)$ and $\mathcal{H}_1 : x_i(n) = s_i(n) + w_i(n)$, where $i=1,2,...,M$. $w_i(n)$ denotes the additive Gaussian noise process with zero mean and variance σ_n^2. In the testing, $s_i(n)$ indicates the band-limited primary signal received at the secondary detector i. It is assumed having a zero mean and a power σ_s^2 similarly to the noise. We also assume that the primary signals $s_i(n)$ and noise $w_i(n)$ received at each secondary detector are independent so as to the validity of the central limit theorem. n is the sample denote from 1 to N(total number in the sample interval).

In spectrum sensing, energy detector is an easy realized detector for detection of unknown signals. It's high-powered in the high signal-noise-ratio(SNR) environment especialy[10]. Central/Non-central Chi-square distribution is valid in the case[10]. In a low SNR scene, however, the central limit theorem becomes incontestable because it need a large samples number-N, when the Central/Non- central Chi-square distribution goes into ill-suited. It collect the test statistics and compare it with a threshold to decide whether the primary signal is present or not. The test statistic is given by $T(Y_i) = 1/N \sum_{n=1}^{N} y^2(n)$ .Where N is the samples number in the detection. $T(Y_i)$ is compared with threshold λ_i to make the final decision. λ_i is set to meet the target probability of false alarm P_F according to the noise power. The probability of detection P_D can be also identified. When there is no uncertainty and the noise variance is completely known, the central limit theorem gives the following approximations:

$$\begin{cases} T(Y)|\mathcal{H}_0 \sim \mathcal{N}(\sigma_n^2, \frac{1}{N} 2\sigma_n^4) \\ T(Y)|\mathcal{H}_1 \sim \mathcal{N}(\sigma_s^2 + \sigma_n^2, \frac{1}{N} 2(\sigma_s^2 + \sigma_n^2)^2) \end{cases} . \tag{1}$$

Using these approximations in (1), the false alarm probability (P_F) and the detection probability(P_D) under the threshold λ_i are given as:

$$\begin{aligned} P_{Fi} &= Prob(T(Y) > \lambda_i \,|\mathcal{H}_0) \\ &= Q\left(\frac{\lambda_i - \sigma_n^2}{\sqrt{2/N}\,\sigma_n^2}\right) \end{aligned} . \tag{2}$$

$$\begin{aligned} P_{Di} &= Prob(T(Y) > \lambda_i \,|\mathcal{H}_1) \\ &= Q\left(\frac{\lambda_i - (\sigma_n^2 + \sigma_n^2)}{\sqrt{2/N}\,(\sigma_n^2 + \sigma_n^2)}\right) \end{aligned} . \tag{3}$$

$Q(\bullet)$ is the standard Gaussian complementry difference function (CDF):

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} \exp(-\frac{x^2}{2}) dx$$

It's easy to educe that samples number $N \propto 1/SNR^2$, then the signals could be detected at arbitrarily low SNR by increasing the sensing time based on the known noise power σ_n^2. However, it is almost impossible to have an exact knowledge of the noise power in the real-world. As the thermal noise or interference from remote transmitters keep changing, the noise power uncertainty in detection may varies with time. The performance of detector should be impacted as the equations(4)(5) have to be modified. It should be considered seriously, especialy as the SNR is very low.

2.2 Noise Uncertainty Model

R. Tandra and A. Sahai have studied the "SNRwall" based on the noise uncertaity bound $[1/\rho, \rho]$ profoundly[5,6]. But they did not give out the uncertaity distribution. A. Sonnenschein and P. M. Fishman[7] gave a moderate peak-to-peak uncertainty based on a normal distribution of the noise power difference. Yonghong Zeng etc. analyzed the ROC of the detector assuming the uncertaity distribution as uniform distribution in dB. But it seems to be unreasonable to describe a fluctuated random process. A log-normal distributed uncertaity model is formulated and its impacts to the detector are studied next.

We assume the true noise power in practice is σ^2 , and $\Delta = \sigma^2 - \sigma_n^2$ (in dB) denoting the noise power random fluctuation, which is a new random variable. It's

well-founded to assume the distribution of Δ as a Gaussian normal distribution with a zero mean and a variance σ_Δ^2 based on the statistic theory: $\Delta\big|_{\mathrm{dB}} \sim \mathcal{N}(0, \sigma_\Delta^2)$.

Then we can educe a log-normal distribution of the real noise power value . Assuming $\sigma^2 / \sigma_n^2 = \rho$, then $\rho = 10^{\Delta/10}$, it's easy to get:

$$\rho \sim \mathcal{L}og - \mathcal{N}(0,\ (\frac{\ln 10}{10}\sigma_\Delta)^2) \cdot$$

The probability density function of ρ can be formulated as:

$$f_\rho(x;\sigma_\Delta)=\begin{cases}\dfrac{10}{\ln 10\sqrt{2\pi}x\sigma_\Delta}\exp\left(-\dfrac{50}{(\ln 10\sigma_\Delta)^2}(\ln x)^2\right), & x>0\\ 0, & x<0\end{cases}. \tag{4}$$

Variance σ_Δ^2 depicts the fluctuation scope of the real noise power σ^2 compared with the referential noise power σ_n^2. Next studies are performanced basing on this distribution of the noise uncertainty ratio ρ in diverse variances. It shows more rational than the uniform distribution model[8] evidently.

3 Detection Performance

3.1 Single Detection

By averaging the noise uncertainty factor, we obtain the expected detection probabilities from formulae(2,3,4) as:

$$\hat{P}_{Fi}(\lambda_i^{'}, \sigma_{\Delta,i}) = \int_0^{+\infty} Q\left(\frac{\lambda_i^{'}-\rho}{\sqrt{2/N\rho}}\right) \times \frac{10}{\ln 10\sqrt{2\pi\rho}\sigma_{\Delta,i}}\exp\left(-\frac{50}{(\ln 10\sigma_{\Delta,i})^2}(\ln\rho)^2\right)d\rho \tag{5}$$

$$\hat{P}_{Di}(\lambda_i^{'}, \sigma_{\Delta,i}) = \int_0^{+\infty} Q\left(\frac{\lambda_i^{'}-(\gamma+\rho)}{\sqrt{2/N(\gamma+\rho)}}\right) \times \frac{10}{\ln 10\sqrt{2\pi\rho}\sigma_{\Delta,i}}\exp\left(-\frac{50}{(\ln 10\sigma_{\Delta,i})^2}(\ln\rho)^2\right)d\rho \tag{6}$$

Note that threshold $\lambda_i^{'} = \lambda_i / \sigma_n^2$, signal-noise-ratio $\gamma = \sigma_s^2 / \sigma_n^2$, $\hat{\bullet}$ denotes an expected value. The thresholds $\lambda_i^{'}$ are determined with the desirable probability of false alarm based on the equation (10). It shows that the two probabilitis are functions

of the variance σ_{Δ}^2 after integrating noise fluctuation ratio ρ . Fig.1. illuminates the ROC of the miss-detection probability($P_M = 1 - P_D$) against the false alarm probability(P_F) in several different noise uncertanty scenes with fluctuation standard variances σ_{Δ} . The curve ($\sigma_{\Delta} = 0$) is a referrence with no fluctuation in the Fig. Our

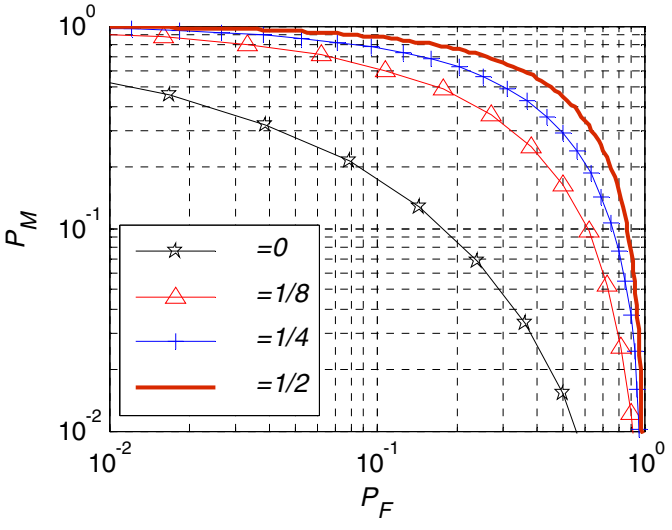


Fig. 1. ROC of EDs in different noise fluctation denoted with diverse σ_{Δ} in dB comparing to the ED in no noise uncertainty with $\sigma_{\Delta} = 0$.

numerical simulation is performed with a low SNR $\gamma = -15dB$ and a large sample number N=10000. It shows that the detector is deteriorated remarkably even in a slight noise power fluctuation, e.g. $\sigma_{\Delta} = 1/8$. The detector almost becomes nonsensical when the fluctuation scope enlarges to $\sigma_{\Delta} = 1$, due to the "SNRwall"[5], though a wealthy sample number already becomes helpless.

3.2 Cooperative Detection

Cooperative communication or cooperative spectrum sensing has been studied deeply for a long time[10]. It has been proved that cooperation could improve sensing reliability greatly. Due to the complexity and susceptivity of the data fusion rules, "AND" or "OR" fusion rules are considered optimum in cooperation. In this paper, we apply the "OR" fusion rules to analyse the performance of cooperative detectors.

In "OR" fusion rules, when at least 1 out of M secondary users detects the presence of primary users, the final decision declares a primary user presence. Therefore, the formulae of PF and PM in the final decision are respectively:

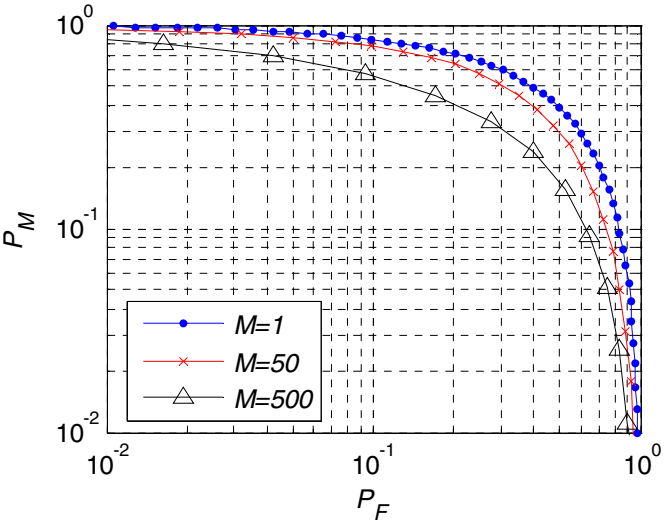


Fig. 2. ROC of cooperation "OR"rules fusion detection in different cooperators denoted with M, and the samples number N=10000, SNR=-15dB, noise uncertainty $\sigma_\Delta = 1/2$

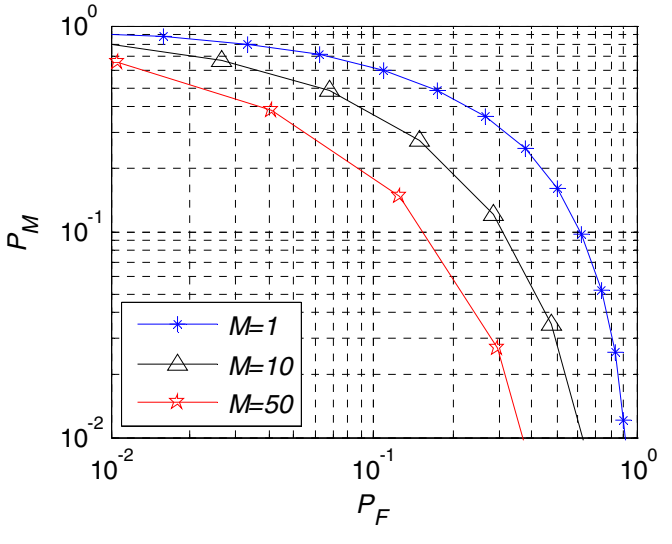


Fig. 3. ROC of cooperation "OR"rules fusion detection in different cooperators denoted with M, and the samples number N=10000, SNR=-15dB, noise uncertainty $\sigma_\Delta = 1/8$

$$\begin{cases} P_F = 1 - \prod_{i=1}^{M} (1 - \widehat{P}_{F,i}) \\ P_M = \prod_{i=1}^{M} (1 - \widehat{P}_{D,i}) \end{cases} . \tag{7}$$

Simply, we assume the M users are in the same noise scene, two defferent σ_Δ uncertainty instances cooperation curves are illuminated in Fig.2 and Fig.3 based on (7). Each Fig. is with 10000 sample number and -15dB SNR and a set of distinct cooperators number denoted in M. We find that decision fusion rules are almost invalid when the noise power fluctuation scope is marked. Fig.2 shows the ROC of different M cooperating sensors with a noise uncertanty variance σ_Δ^2=1/4. It shows that the performance is quite unacceptable though there are 500 cooperators working togather. However, cooperation takes effect when the uncertainty minishes to a certain scope. Fig.3 reveals an acceptable cooperation performance of moderate cooperators when the uncertainty standard variance diminishes to 1/8. We consider it due to the low-grade single detecting performance in a smart noise fluctuation relative to the low SNR.

4 Conclusions and Future Works

Cognitive Users(CU) need to monitor the primary signals continuously during their working time. However, the uncertaitive noise power deteriorates the performance of spectrum sensing greatly, in a low SNR environment. In this paper, we propose a log-normal noise power uncertainty model, and study the impacts of the uncertainty to the detector based on the proposed model. We find that a tiny fluctuation of the noise power can deteriorate the validity of the single ED in low SNR. Applying "OR" fusion rules algorithm of cooperative spectrum sensing, we find that cooperative sensing can improve the performance moderately when the uncertainty standard variance is not too large. However, cooperated ED system is invalid when the uncertainty enlarges to a certain range. Though the correlativities of the SNR and the uncertainty scope have been studied resulting in a "SNRwall" notion[5,6], the correlativity between noise uncertainty distribution characteristics(σ_Δ^2) and SNR or samples number N need further study. A valid and efficient detection system in low SNR situation and arbitrary noise uncertainty should be explored further.

Acknowledgments. This work was supported by state 973 projects (No.2007CB310607), important national science & technology specific projects (2009ZX03003-002, 2009ZX03007-004).

References

1. Haykin, S.: Cognitive radio: Brain-empowered wireless communications. IEEE Journal on Selected Areas in Communications 23(2), 201–220 (2005)

2. Federal Communications Commission, .Notice of Proposed Rulemaking, in the matter of unlicensed operation in the TV broadcast bands (ET Docket No. 04-186) and additional spectrum for unlicensed devices below 900 MHz and in the 3 GHz band (ET Docket No. 02-380), FCC 04-113 (May 2004)
3. http://www.ieee802.org/22/
4. Urkowitz, H.: Energy detection of unkown deterministic signals. Proceedings of the IEEE 55(4), 523–531 (1967)
5. Tandra, R., Sahai, A.: Fundamental limits on detection in low SNR under noise uncertainty. In: 2005 International Conference on Wireless Networks, Communications and Mobile Computing, pp. 464–469 (2005)
6. Tandra, R., Sahai, A.: SNR walls for signal detection. IEEE Journal of Selected Topics in Signal Processing 2, 4–17 (2008)
7. Sonnenschein, A., Fishman, P.M.: Radiometric detection of spreadspectrum signals in noise of uncertain power. IEEE Trans. Aerosp. Electron. Syst. 28, 654–660 (1992)
8. Zeng, Y., Liang, Y.-C., Hoang, A.T., Peh, E.C.Y.: Reliability of Spectrum Sensing Under Noise and Interference Uncertainty. In: Proc. - IEEE Int. Conf. Commun. Workshops, ICC 2009 (2009)
9. Lin, W., Zhang, Q.: A design of energy detector in cognitive radio under noise uncertainty. In: ICCS 2008, pp. 213–217 (November 2008)
10. Letaief, K.B., Zhang, W.: Cooperative Communications for Cognitive Radio Networks. Proceedings of the IEEE 97(5) (May 2009)

A Realtime Human-Computer Ensemble System: Formal Representation and Experiments for Expressive Performance

Tetsuya Mizutani[1], Shigeru Igarashi[1], Tatsuo Suzuki[2,⋆], Yasuwo Ikeda[3], and Masayuki Shio[4]

[1] Department of Computer Science, University of Tsukuba, Tsukuba, Japan
{mizutani,igarashi}@cs.tsukuba.ac.jp
[2] Master's Program in Computer Science, University of Tsukuba, Tsukuba, Japan
tatsuo@jewel.ocn.ne.jp
[3] Mejiro University, Tokyo, Japan
yikeda@krd.biglobe.ne.jp
[4] College of Community Development, Tokiwa University, Mito, Japan
shio@tokiwa.ac.jp

Abstract. A human-computer ensemble system is one of time concerned cooperative systems, which performs secondo of an ensemble played by a computer-controlled piano cooperating with primo played by a human performer. For creating expressive performance, a rehearsal program is adopted to this system. By the rehearsal program, the system learns the tendency of the expression that the human performer thinks and/or plans. To do so, the program records his/her performance of solo and calculates the tendency, which depends on not only the performer but also the composition of music itself. Hence, it is necessary for the expressive performance to analyze the score of the composition and to experiment dependent on it.

This system is an example of intelligent realtime programs appropriate for formal verification and analysis. $N\Sigma$-labeled calculus is a formal system for verification for such time-concerned programs.

In this paper, a logical specification and experimental results for an expressive ensemble system will be introduced.

Keywords: Artificial Intelligence, Musical Informatics, Ensemble System, Expression of Music, Formal Specification.

1 Introduction

A human-computer ensemble system [11,14,15,17] is one of time concerned cooperative systems, which performs secondo of an ensemble played by a computer-controlled piano cooperating with primo played by a human performer. It is an example of intelligent realtime programs appropriate for formal verification and analysis. *$N\Sigma$-labeled calculus* [8,10,12,13] is a formal system for verification

⋆ Current Affiliation: Nikko Systems Solutions, Ltd., Yokohama, Japan.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 256–265, 2010.

for such *time-concerned* programs, which are meant intelligent programs of machines, or of humans, to treat matters not only mathematically time-dependent but also distinctively featured by subtle and sophisticated sensitivity, interest or human concern in exact time or timing, such as traffic control, ensemble in music, etc.

For creating expressive performance by the ensemble system, we adopt the *rehearsal* program. By the rehearsal program, the system learns the tendency of the expression that the human performer thinks and/or plans. To do so, the program records his/her performance of solo and calculates the tendency. The tendency of expression depends not only on the performer but also on the composition of music itself. Hence, it is necessary for the expressive performance to analyze the score of the composition and to experiment dependent on it.

In this paper, a logical specification and experimental results for an expressive ensemble system will be described.

2 Ensemble System with "Rehearsal"

The ensemble system consists of two programs; the *rehearsal* and the *performance* ones.

The aim of the system is to perform *secondo* (the second part) of an ensemble to drive a MIDI acoustic piano by the performance program cooperating with performing *primo* (the first part) by a human pianist or other musician playing another MIDI instrument. Those two MIDI instruments are connected by a computer. MIDI (Musical Instrument Digital Interface) is a standard protocol for electronic musical instruments and computers to communicate, control, and synchronize with each other. A MIDI acoustic piano is an acoustic grand piano that can be played with MIDI codes. There is a 0.5[s] delay between the moment when each MIDI code as the input of the the piano and that when the corresponding key moves and actually sounds, since the actuator must adjust the timing of each group of notes which sound simultaneously. Hence, the program must output each MIDI code 0.5[s] before the moment when the actual performance is expected.

The program has the whole score data of ensemble. There are special notes called *reference notes* to measure the performance speed (or 'local tempo'), usually at the head of each or every other bar, depending on the structure and expected tempo of the score.

Before the cooperating performance, the rehearsal program records solo primo performed by the human performer. From this performance, the program calculates the *rehearsal data* which expresses the timing of performance of the whole reference notes, and *schedule* of performance of secondo. During the cooperating performance, the program modifies and updates the schedule in realtime from the actual performance of primo using the score data (including the reference notes), the rehearsal data and the schedule.

3 Specification of the Ensemble System

The specifications of the ensemble system, consisting of the rehearsal and the performance programs, are represented along with [14,15]. The specification of the former is represented rather informally, since it is simple, and moreover, not realtime processing, i.e. only the input/output relationship. Thus, it is easy to represent this specification formally. In contrast, the latter must be described by $N\Sigma$-labeled calculus since it is a rather complicated time-concerned program involving a kind of intellectual processing.

In these representations, at least four functions denoting below are important to design the system, where ω is the ordinal number of the natural numbers and $\omega + 1$ is ω with ∞.

- $\varphi : \omega \to \omega + 1$: each note of primo sent onto its "note-on" time actually played by a human performer, or ∞ if it is not played,
- $\phi : \omega \to \omega + 1$: the time when a reference note of primo performed,
- $\Psi : \omega \to (\omega + 1) \times 2$: sending each MIDI code onto a pair ⟨the predicted time of issuing codes, "note-on"/"note-off"⟩ of secondo, the former of which is expressed by $\Psi(m).\texttt{time}$ and the latter $\Psi(m).\texttt{play}$.
- $\psi : \omega \to (\omega + 1) \times 2$: the time of each note in secondo part sent onto ⟨"note-on" time, "note-off" time⟩, the former and the latter of which is expressed by $(\psi(x))_0$ and $(\psi(x))_1$, respectively, so as to issue MIDI codes either to depress or to release the key corresponding to the m-th note in secondo at these time, or never ($=\infty$).

It is essential to "predict" and "correct" for playing the secondo performance Ψ, or its actual output ψ, along with the that of primo φ, especially that of the reference notes ϕ. Hence, the correspondence between ϕ and ψ is very important for both design and verification of the system.

Additionally, the following variables and functions as indexes are used in the specification.

- $l \in \omega$ expresses the order of each primo note in the score.
- j: $\omega \to \omega$ is a function designating that n-th reference note is the $j(n)$-th note at the original order in the score.
- $x \in \omega$ denotes the order of each secondo note.
- u and v: $\omega \to \omega$ are functions that maps each second note onto the corresponding note-on and note-off MIDI codes, respectively.

3.1 Specification of the Rehearsal Program

The input of this program is a primo performance by a human performer and the output is the correspondence secondo data as the rehearsal data of the input of the performance program with the same or very closed expression to the primo performance.

Though the measurement of the timing of the input of each primo note is processed in realtime, the calculation and creation of the corresponding secondo data is not in realtime.

Table 1. Axiom tableau for the ensemble program

index	*condition*	*action*	*tense*	*label*
1.1	$\Psi(\mathtt{m}).\mathtt{time}$ $\leq \mathtt{Timer} + 500[\mathrm{ms}]$	$c_1 < \alpha = a_2 \leq d_1$	$1^{\boldsymbol{st}} a_1,\ t$	$*,\ \boldsymbol{S}$
1.2	**otherwise**	$c_1 < \alpha = \mathit{GetState} \leq d_1$		
2.1	$\mathtt{m} = u(x)$, $\Psi(\mathtt{m}).\mathtt{time}$ $\leq \mathtt{Timer} + 400[\mathrm{ms}]$, $\Psi(\mathtt{m}).\mathtt{play} = \mathrm{on}$	$c_2 < \alpha = a_4 \leq d_2$, $(\psi(x))_0 = \infty[\mathrm{ms}]\ @\alpha$	$1^{\boldsymbol{st}} a_2,\ t$	$*,\ \boldsymbol{S}$
2.2	**otherwise**	$c_2 < \alpha = \mathit{MIDIOut} \leq d_2$		
3.1	$\mathtt{m} = u(x)$	$(\psi(x))_0 = \mathtt{Timer} + 500[\mathrm{ms}]@\alpha$	$1^{\boldsymbol{st}}\ \mathit{MIDIOut},\ t$	$*,\ \boldsymbol{S},\ \boldsymbol{P},\ \boldsymbol{L}$
3.2	$\mathtt{m} = v(x)$	$(\psi(x))_1 = \mathtt{Timer} + 500[\mathrm{ms}]@\alpha$		
3.3	-	$c_3 < \alpha = a_4 \leq d_3$		
4	$\mathtt{m} = x$	$\alpha = \mathit{GetState} < \varepsilon$, $\mathtt{m} = x + 1\ @\alpha$	$1^{\boldsymbol{st}} a_4,\ t$	$*,\ \boldsymbol{S}$
5.1	$G(l,\ p)$,**global**	$c_5 < \alpha = a_6 \leq d_5$	$(p+1)^{\boldsymbol{st}}\ \mathit{GetState}$	$*,\ \boldsymbol{S},\ \boldsymbol{P},\ \boldsymbol{L}$
5.2	**otherwise**, **global**	$\alpha = a_1 < \varepsilon$	$(p+1)^{\boldsymbol{st}}\ \mathit{GetState}$	
6.1	$\mathit{Match}(\mathtt{n},\ \varphi(l),\ \phi,\ j)$, $\mathtt{Gap} = x$	$(\ \mathtt{Rate} = \mathit{Rate}(\mathtt{n}-1,\ x),\ \mathtt{Gap} = \mathit{Gap}(\mathtt{n}),\ l = j(\mathtt{n}))@\alpha$	$1^{\boldsymbol{st}} a_6,\ t$	$*,\ \boldsymbol{S}$
6.2	-	$c_6 < \alpha = a_1 \leq d_6$		
7.1		$\mathtt{Timer} = [\frac{U-\theta}{4\times 10^6}]$	t	$*,\ \boldsymbol{S}$
7.2		$\alpha = a_1 = 1,\ (\mathtt{m} = 1, \mathtt{Gap} = 0,\ \mathtt{Rate} = 1)\ @\alpha$		
7.3		$U = t + \theta$	t	
8		$\beta_{\boldsymbol{l\text{-}th}} = \varphi(l)$		$*,\ \boldsymbol{P},\ \boldsymbol{L}$

$\boldsymbol{S}$: The ensemble program, $\boldsymbol{P}$: The person pianist, $\boldsymbol{L}$: Listeners.

The detailed specification is as follows. For the note-on time $\varphi_s(x)$ of the x-th note of the secondo is calculated

- if there is a primo note l that must be performed simultaneously with x on the score, then $\varphi_s(x) = \varphi(l)$. Especially, if l is a reference note $j(n)$, then $\varphi_s(x) = \phi(n)$,
- otherwise, let l_1 and l_2 be the primo note just before and after x on the score, respectively. Let b_1 be the metric length (the length on the score) between l_1 and l_2, and b_2 be that between l_1 and x. Then, $\varphi_s(x) = \varphi(l_1) + \frac{b_2}{b_1}(\varphi(l_2) - \varphi(l_1))$.

3.2 Specification of the Performance Program

A formal specification of the performance program of the system represented in $N\Sigma$-calculus has been introduced in [14]. Table 1 is the specification represented by a *axiom tableau* [9,10].

Explanation of variables and functions other than those described in the above are as follows.

- c_i and d_i $(1 \leq i \leq 6)$ are the lower and upper bounds of computation and performance time of each program block, respectively. ε is also an upper bound that is only a few nanoseconds, i.e. a few steps of machine instruction.
- `n` is a program variable expressing the order in the score of the reference note just processing.
- `m` is a program variable denoting the order of predicted output of MIDI code "note-on"/"note-off".
- `Timer` is a program variable representing the timer counting up every 4 [ms], which is the resolution of the MIDI acoustic piano.

- `Rate` and `Gap` are program variables expressing the ratio of primo tempo and that of the rehearsal data, and the gap or difference between the performance time of a primo note and that of the corresponding rehearsal note, respectively.
- $Rate: \ \omega \to \omega \to \omega$ and $Gap: \ \omega \to \omega$ are functions expressing the ratio between the tempo of primo performance and that of the rehearsal data, and the gap or difference between the performance time of a primo note and that of the corresponding rehearsal note, respectively.
- α and β are *spurs*, served as the generalizations of schedulers [7], for the performance program and for the human performer.
- a_1, a_2, *MIDIOut*, a_4, *GetState* and a_6 are *program labels*.

Each program axiom is explained briefly.

- Axiom 1 (indexed by 1.1 and 1.2 in Table 1): When the program label a_1 arises, expressed by "$1^{\boldsymbol{st}} a_1, t$" in the tense cell, one of the following actions 1.1 or 1.2 is performed.
  - 1.1: If $\Psi(\mathtt{m}).\mathtt{time} \leq \mathtt{Timer} + 500$[ms] in the condition cell, then the next action is to rise a_2, performed between c_1 [ns] and d_1 [ns], expressed in the action cell.
  - 1.2: Otherwise, the control goes to *GetState* block between c_1 [ns] and d_1 [ns] to get a new MIDI input from the performer.
- Axiom 2: When a_2 arises,
  - 2.1: if The next MIDI code $\Psi(\mathtt{m}).\mathtt{play}$ is on, if $\mathtt{m} = u(x)$, and if $\Psi(\mathtt{m}).\mathtt{time} \leq \mathtt{Timer} + 400$[ms], then the note-on MIDI output is not played, i.e., $(\psi(x))_0 = \infty$, which operation is performed between c_2 [ns] and d_2 [ns], and
  - 2.2: otherwise, the control goes to *MIDIOut* block between c_2 [ns] and d_2 [ns] to output the MIDI code.
- Axiom 3: When *MIDIOut* arises,
  - 3.1: If $\mathtt{m} = u(x)$, then the note-on time $(\psi(x))_0$ is $\mathtt{Timer} + 500$[ms].
  - 3.2: If $\mathtt{m} = v(x)$, i.e. the next output is note-off, then the note-off time $(\psi(x))_0$ is $\mathtt{Timer} + 500$[ms].
  - 3.3: Each of the below operation is performed between c_3 [ns] and d_3 [ns].
- Axiom 4: When a_4 arises, `m` counts up 1 and go to *GetState* within a few (ε) nanoseconds.
- Axiom 5: When the program controller reached *GetState* $(p+1)$-st times from the period when the system started, represented by $(p+1)^{\boldsymbol{st}}$ *GetState* in the tense cell,
  - 5.1: if there is an input code at the input MIDI port, expressed by $G(l, \ p)$, then it is recognized as the l-th primo note and the controller goes to a_6 between c_5 [ns] and d_5 [ns], where $G(l, \ p)$ is an abbreviation of $p^{\boldsymbol{th}}$ *GetState* $< \varphi(l) \leq (p+1)^{\boldsymbol{st}}$ *GetState*, and,
  - 5.2: otherwise, the controller goes to a_1 between c_5 [ns] and d_5 [ns].
- Axiom 6: When a_6 arises,

  - 6.1: if the input code matches the l-th primo note successfully ($Match(\mathtt{n},\ \varphi(l),\ \phi,\ j)$), then the value of `Rate` and `Gap` are calculated from the these old values, and the value of `n` becomes to satisfy $l = j(\mathtt{n})$, and
  - 6.2: the controller goes to a_1 either the matching succeeds or not.
- Axiom 7 is for `Timer` and the initial conditions.
  - 7.1 is for `Timer` counting up every 4[ms].
  - 7.2 represents the initial values. Initially, the program label is a_1, and the values of `m`, `Gap` and `Rate` are 0, 1, 1, respectively.
  - The free variable U represents the global clock and θ is a free variable indicates the initialization time of the system.
- Axiom 8 represents the behavior of the human performer. The l-th the note-on is at $\varphi(l)$.

Essential properties for the system can be verified and analyzed by $N\Sigma$-labeled calculus. Especially, the following property has been be proved in the calculus in [14,15].

Theorem 1. $(\psi(x))_1 < \infty$, *every* $x < M$,
i.e. a note-off signal is issued for the x*-th secondo note for all* $x < M$*, where* M *is the number of secondo notes in the score, provided that every matching of the reference notes from the first to the end is successful, i.e.*

$$\forall n \leq N(Match(n,\ \varphi(j(n)),\ \psi,\ j))$$

holds, where N *is the number of the reference notes in the score.*

4 Experiments for Expressive Performance

The ensemble system has been implemented based on the specification described in the previous section. In this system, the matching program between the stream of input MIDI codes and the score data of the primo has been referred to the best match algorithm in [4] that is used the dynamic programming technic.

By many actual experiments with a professional human performer performed solo part, ensemble performances at a certain level have been obtained. However, we have also obtained the result that there are many points to modify the system for the further expressive performance.

Figure 1 is the first page of the score for ensemble of flute and piano of Humoreske Op. 101-7 by A. Dvořák [2], which is used in the actual experiments for the system. It has 16 bars and there is a repeat mark after the 8th bars, then it is performed as 24-bar length.

The tempo and rhythm must change at the 17th bar (the 9th bar after the repeat). In general, the performance of between 1st and 16th bar is faster than that after 17th bar. However, there is no clue of the expression from the rehearsal, since there is only one note of the primo in the eight bar. Thus, the prediction of the tempo only from the rehearsal data is very difficult.

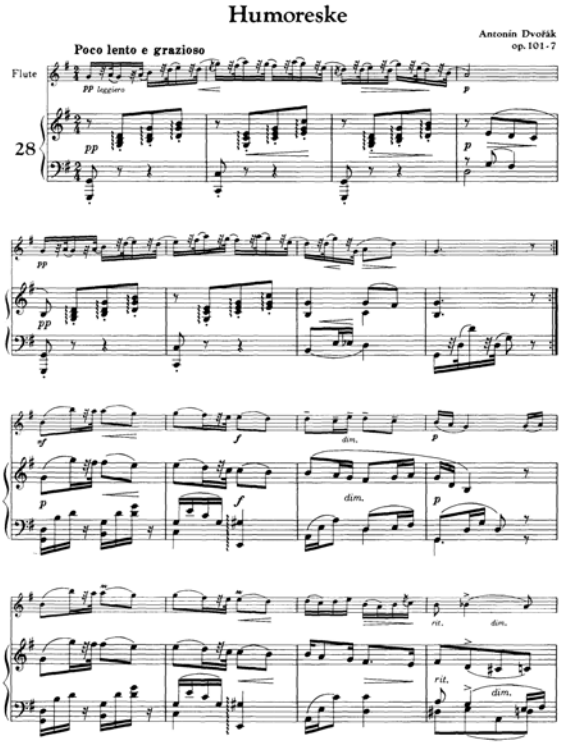


Fig. 1. Humoreske Op. 101-7

Several solutions for such problems have been considered, e.g. by the HMM (Hidden Markov Model), the probabilistic or other statistical methods. Unfortunately, these methods do not suitable in this case since they are for the solo performance or for precise accompaniments, i.e. there are few gap of time between each corresponding play of primo and secondo notes. Thus, we must consider and experiment other solutions.

The following 4 candidates of solution have been considered and experimented.

(a) *No* special methods are used.
Namely, it follows the original specification.

(b) The data from the performance of the secondo of the 8th bar are added for the rehearsal data.
To create the rehearsal data, the human performer first plays the primo solo, and next, does the secondo of the 8th bar and the first note of the 9th bar. The rehearsal program calculates the expression of the secondo using the secondo play data by the human performer.
It is rather reliable method since the expression is given from the human performer directly, but, of course, his/her load increases.

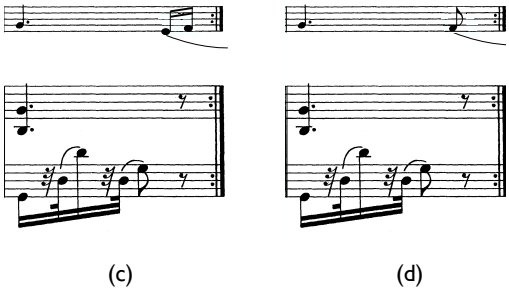


Fig. 2. Methods (c) and (d): (c) is added two 16th notes at the end beat of the 8th bar, while (d) one 8th notes at the end beat of the 8th bar.

(c) Two 16th notes for the *Auftakt* (the *upbeat*) at the end beat of the 8th bar are added for the rehearsal.
Figure 2(c) is the score of the 8th bar that has two added 16th notes as the Auftakt.
For creating the rehearsal data, the human performer plays the primo including these notes. The rehearsal program calculates the secondo data using the timing of the adding notes. Note that the human performer does note play these notes on the cooperative performance. It seems suitable since the rehearsal program can calculate the timing of the 9th bar after the repeat easily, and moreover, the load of the human rarely increases.

(d) One 8th note for the Auftakt at the end of the 8th bar are added for the rehearsal (Figure 2(d)).
It is similar to the method (c).

Through many experiments, as shown in Table 2, the result is obtained that one of the best solutions is (b).

The experiments of the methods (c) and (d) end in failure, since the human performer performed the 8th bar with the Auftakt note(s) about 400 ms (c) or 250 ms (d) longer than that without such a note when she played at the rehearsal, which are not, of course, the intended expression of the performance. It follows one hypothesis that performance of professional or experimental performers remarkably different between the same scores but one has a rest and the other has a correspondence note, although the they are unconscious.

5 Related Works

Many complicated problems have been verified and analyzed by the calculus, e.g. an existing trains accident [9], a vehicle controlling system [10], and an existing airplanes near miss accident [12,13].

Table 2. Result of the Experiments

	(a)	(b)	(c)	(d)
(1)	74.4	**47.2**	59.7	67.8
(2)	74.8	**58.0**	77.3	74.0
(3)	5900.1	**3987.9**	9485.6	9422.4

(1) Average of the absolute value of the gap between the notes of primo and the corresponding secondo ones [ms],
(2) average of the absolute value of the gap between the *reference* notes of primo and the corresponding secondo ones [ms],
(3) the variance;

There are many reasoning systems for temporal or time-related logic of knowledge and belief of human [1,5,6], etc. And also, airplane controlling systems as well as vehicle controlling systems are formally treated in [3]. All of these systems use *model checking* method, while $N\Sigma$-labeled calculus does the *proof-theoretic* method on the *natural number theory*, which is one of the remarkable feature of this calculus.

And also, there are many related works for automatic ensemble systems. Most of them, e.g. [16,18], etc. use HMM or other statistical methods to predict future rhythm and tempo of performance by human performer. Compared with them, this system uses *rehearsal* data created by the rehearsal of solo by a human performer to learn the expression pattern, while the cooperating program only calculates the tempo from the "history" of the performance.

6 Conclusion

A formal specification of the ensemble system is introduced by the formal system $N\Sigma$-labeled calculus. And also, it has been successfully implemented based on this specification.

Though there is only one experimental validation and there is no logical verification that the solution described in Section 4 is the *best*, the authors think that it must be *better* than most of other candidates since it uses performances by a human performer directly for creating the correspondence expression.

One of our future works is that the uniform method for learning and creating expression. Though there are many studies for expression of music, we think that it is very difficult to adopt the ensemble system since the expression is depends both on the score and the human performer.

Acknowledgment

The authors would like to thank Ms. M. Matsuzaki (Pf.) who attended the experiment, gave them many valuable data through playing music with the system, and gave many useful advice for music.

References

1. Curzon, P., Ruksenas, R., Blandford, A.: An Approach to Formal Verification of Human-Computer Interaction. Formal Aspect of Computing 19, 513–550 (2007)
2. Eto, Y. (ed.): Flute Famous Collections 31, p. 112. Doremi Music Publishing (1986)
3. Damm, W., Hungar, H., Olderog, E.-R.: Verification of Cooperating Traffic Agents. International Journal of Control 79, 395–421 (2006)
4. Dannenberg, R.B.: An On-Line Algorithm for Real-Time Accompaniment. In: Proceedings of the 1984 International Computer Music Conference, pp. 193–198 (1985)
5. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning About Knowledge. The MIT Press, Cambridge (1995)
6. Halpern, J.Y., Vardi, M.Y.: The Complexity of Reasoning about Knowledge and Time. I. Lower Bounds. Journal of Computer and System Sciences 38, 195–237 (1989)
7. Igarashi, S.: The ν-Conversion and an Analytic Semantics. Information Processing 83, 769–774 (1983)
8. Ikeda, Y., Mizutani, T., Shio, M.: Formal System and Semantics of $N\Sigma$-labeled Calculus. In: The 2009 International Conference on Artificial Intelligence and Computational Intelligence, vol. III, pp. 270–274. IEEE, Los Alamitos (2009)
9. Mizutani, T., Igarashi, S., Ikeda, Y., Shio, M.: Labeled @-Calculus: Formalism for Time-Concerned Human Factors. In: Calmet, J., Ida, T., Wang, D. (eds.) AISC 2006. LNCS (LNAI), vol. 4120, pp. 25–39. Springer, Heidelberg (2006)
10. Mizutani, T., Igarashi, S., Shio, M., Ikeda, Y.: Human Factors in Continuous Time-Concerned Cooperative Systems Represented by $N\Sigma$-labeled Calculus. Frontiers of Computer Science in China 2, 22–28 (2008)
11. Mizutani, T., Suzuki, T., Shio, M., Ikeda, Y.: Formal Specification and Experiments of an Expressive Human-Computer Ensemble System with Rehearsal. In: TASE 2009, The 3rd IEEE International Symposium on Theoretical Aspects of Software Engineering, pp. 303–304 (2009)
12. Mizutani, T., Igarashi, S., Shio, M., Ikeda., Y.: Analysis of an Airplane Accident in $N\Sigma$-labeled Calculus. In: TASE 2009, The 3rd IEEE International Symposium on Theoretical Aspects of Software Engineering, pp. 311–312 (2009)
13. Mizutani, T., Igarashi, S., Shio, M., Ikeda, Y.: Formal Analysis of an Airplane Accident in $N\Sigma$-labeled Calculus. In: Deng, H., Wang, L., Wang, F.L., Lei, J. (eds.) AICI 2009. LNCS, vol. 5855, pp. 469–478. Springer, Heidelberg (2009)
14. Mizutani, T., Igarashi, S., Shio, M., Ikeda, Y.: Labeled Calculi Applied to Verification and Analysis of Time-Concerned Programs I, TENSOR, N. S. (to appear)
15. Mizutani, T., Igarashi, S., Shio, M. and Ikeda, Y.: Labeled Calculi Applied Verification and Analysis of Time-Concerned Programs II, TENSOR, N. S. (to appear)
16. Raphael, C.: Automated Rhythm Transcription. In: ISMIR 2001: International Symposium on Music Information Retrieval, pp. 99–107 (2001)
17. Suzuki, T.: Creation and Analysis of Expression on an Ensemble System, Master's Thesis, Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba (2009) (in Japanese)
18. Takeda, H., Nishimoto, T., Sagayama, S.: Joint Estimation of Rhythm and Tempo of Polyphonic MIDI Performance Using Tempo Curve and Hidden Markov Models. Journal of Information Processing 48, 237–247 (2007) (in Japanese)

A New Smooth Support Vector Machine

Jinjin Liang[1] and De Wu[2]

[1] Department of Mathematical Sciences, Xi'an Shiyou Universtiy,
Shaanxi Xi'an, China
myonlyonly@126.com
[2] Department of Computer Sciences, Xidian Universtiy,
Shaanxi Xi'an, China
wood1979@163.com

Abstract. A new Smooth Support Vector Machine (SSVM) is proposed and is called NSSVM for short. Different from traditional SSVM that treats perturbation formulation of SVM, NSSVM treats standard 2-norm error soft margin SVM. Different from traditional SSVM that uses the 2-norm of the Lagrangian multipliers vector to roughly substitute that of the weight of the separating hyperplane, which makes the obtained smooth model unequal to the primal program; NSSVM takes into account the connotative relation between the primal and dual program to transform the original program to a new smooth one. Numerical experiments on several UCI datasets demonstrate that NSSVM has higher precisions than existing methods.

Keywords: Smooth Support Vector Machine; 2-norm error soft margin SVM; connotative relation; primal and dual program.

1 Introduction

Support Vector Machine (SVM), recently developed by V. Vapnik and his co-workers, has been a promising method for data classification and regression which is based on statistical learning theory and dual program [1,2]. SVM is reckoned as a new machine learning method for the small-sample sets and is based on structural risk minimization principle [3, 4], which minimizes an upper bound on the generalization error. When points are separable, we figure out the hard margin classifiers that completely separate the two classes which maximize their margin; but sometimes points are non-separable, as is usually the case, so soft margin classifiers are introduced which allow erroneous for misclassified points. The commonly used methods are 1-norm and 2-norm error soft margin SVM. The former is called box constrained SVM, since the Lagrangian multipliers vector satisfies $0 \le \alpha \le C$; and the latter is called diagonal weighted SVM, since the Lagrangian multipliers vector satisfies $0 \le \alpha = C\xi$.

This paper focuses on making use of smoothing techniques, which has been extensively used for solving important mathematical programming problems [5, 6] to

* Supported by National Science Foundation of China under Grant No. 60574075, 60674108.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 266–272, 2010.

obtain a fast and iterative algorithm. Reference [7] first applies smoothing techniques to one specified reformulation of SVM by appending an additional bias term $\gamma^2/2$ to the standard 2-norm error soft margin SVM, which technique do cause changes of accuracies [8] and obtain SSVM. Based on SSVM, diverse kinds of smooth methods have been proposed [9,10,11,12]: these methods aim at designing new smooth functions, or generalizing them to the regression case; but none of them prove the equivalence between the obtained models and the original program. In fact, the 2-norm of the error is roughly used to substitute that of the weight of the hyperplane for all the above mentioned methods, which makes the obtained smooth models unequal to the primal program.

This paper investigates a new Smooth Support Vector Machine (NSSVM), which overcomes the above mentioned disadvantages. The problem treated is the standard 2-norm error soft margin SVM. We begin from the linear case to obtain the unconstrained optimization problem; then we take into account the connotative relation between the primal and dual program to generalize it to the kernel case. Upon obtaining the smooth formulations, we apply Newton algorithm to solve the optimal solution. Various numerical experiments on UCI dataset demonstrate its applicability.

2 New Linear Smooth SVM

Denote T as the training set $T=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}(x_i\in X=R^n)$, and denote L as the label set $L=\{y_1,y_2,...,y_m\}(y_i\in\{1,-1\})$. Training standard 2-norm error soft margin SVM equals solving the following program:

$$\begin{aligned}&\min_{(w,\gamma,\xi)\in R^{n+1+m}} \tfrac{1}{2}w^Tw+\tfrac{v}{2}(\sum_{i=1}^{l}\xi_i^2)\\ &s.t.\ \ y_i(w^Tx_i+\gamma)\geq 1-\xi_i, i=1,...,l\end{aligned}. \tag{1}$$

Here w is the normal to the separating planes, γ is the offset and ξ_i is the error of the i-th training sample, and the linear separating hyperplane is $w^Tx-\gamma=0$. Denote the training set and the label set in the matrix form as $A_{m\times n}$ and $D_{m\times m}=Diag(L)$, then training program (1) equals to (2):

$$\begin{aligned}&\min_{(w,\gamma,\xi)\in R^{n+1+m}} \tfrac{1}{2}w^Tw+\tfrac{v}{2}\xi^T\xi\\ &s.t.\qquad D(Aw-e\gamma)\geq e-\xi\\ &\qquad\qquad \xi\geq 0\end{aligned}. \tag{2}$$

Write the error as $\xi=\max(e-D(Aw-e\gamma),0)=(e-D(Aw-e\gamma))_+$. Using this equation, we can convert (2) into its equivalent unconstrained formulation.

$$\min_{(w,\gamma)\in R^{n+m}} \tfrac{1}{2}w^Tw+\|(e-D(Aw-e\gamma))_+\|^2 . \tag{3}$$

It is easy to prove that this is a convex program, but since the objective function is not differential which precludes the use of a fast Newton method. Define an entropy penalty function as (4), we have two lemmas.

$$P_\beta(t)=t+\beta^{-1}\ln(1+\exp(-\beta t)) \ . \tag{4}$$

Lemma 1. $P_\beta(t)$ is a strict convex function.

Lemma 2. [7] For $x\in R$ and $|x|<\rho : P_\beta(x)^2-(x_+)^2 \le (\frac{\ln 2}{\beta})^2+\frac{2\rho}{\beta}\ln 2$, where $P_\beta(x)$ is defined as in (8) with smoothing parameter $\beta>0$.

Corollary 1. Entropy penalty function $P_\beta(t)=t+\beta^{-1}\ln(1+\exp(-\beta t))$ converges to the plus function $t_+=\max\{0,t\}$, where $\beta>0$ is the smooth parameter.

Then we obtain the smooth model of the 2-norm error soft margin SVM as follows.

$$\min_{(w,\xi)\in R^{n+m}} \tfrac{1}{2}w^T w+\tfrac{v}{2}\| P_\beta(e-D(Aw-e\gamma))\|^2 \ . \tag{5}$$

3 New Kernel Smooth SVM

In the kernel space, suppose the nonlinear map $\phi(x)$ is used to map the original data to a high dimensional one. The optimization program is as follows.

$$\begin{aligned}&\min_{(w,\gamma,\xi)\in R^{n+1+m}} \tfrac{1}{2}w^T w+\tfrac{v}{2}(\sum_{i=1}^{l}\xi_i^2)\\ &s.t.\quad y_i(w^T\phi(x_i)+\gamma)\ge 1-\xi_i, i=1,...,l\end{aligned} \ . \tag{6}$$

Similar as the linear smooth case, we aim at deriving the corresponding smooth model in the kernel case for 2-norm error soft margin SVM. Taking into account the connotative relation between the primal program and the dual one, we transform the original program to one smooth model and prove the equivalence between the obtained model and the original one.

It is known that, at the optimal solution, w is a linear combination of training data:

$$w=\sum_{i=1}^{l} y_i\alpha_i\phi(x_i) \ . \tag{7}$$

Making use of the above relation, we have the following two equations

$$y_i w^T\phi(x_i)=\sum_{j=1}^{l} y_i y_j\alpha_j\phi(x_j)^T\phi(x_i)=\sum_{j=1}^{l} Q_{ij}\alpha_j=(Q\alpha)_i \ . \tag{8}$$

$$w^T w=\sum_{i=1}^{l} y_i\alpha_i\phi(x_i)^T w=\sum_{i=1}^{l}\alpha_i(Q\alpha)_i=\alpha^T Q\alpha \ . \tag{9}$$

So (6) can be expressed as follows:

$$\begin{aligned} &\min_{(\alpha,\gamma,\xi)\in R^{m+1+m}} F(\alpha,\gamma,\xi)=\tfrac{1}{2}\alpha^T Q\alpha+\tfrac{v}{2}\xi^T\xi \\ &s.t.\quad Q\alpha+\gamma Y\geq e-\xi \end{aligned} \quad . \tag{10}$$

Though (10) is different from (6), we will prove that for any optimal α^* of (10), $w^*=\sum_{i=1}^{l} y_i\alpha_i^*\phi(x_i)$ is an optimal solution of (6).

Theorem 1. Denote $w=\sum_{i=1}^{l} y_i\alpha_i\phi(x_i)$, suppose (α^*,b^*,ξ^*) is the optimal solution of (10), and then (w^*,b^*,ξ^*) is the optimal solution of (6).

Proof: Equations (8) and (9) show that: if and only if (w^*,b^*,ξ^*) and (α^*,b^*,ξ^*) are respectively primal and dual optimal solutions, then (α^*,b^*,ξ^*) and (w^*,b^*,ξ^*) are respectively feasible for (10) and (6).

Now we prove this statement that (w^*,b^*,ξ^*) is the optimal solution of (6) can be inferred from the fact that (α^*,b^*,ξ^*) is the optimal solution of (10).

Using equation (7), we have (11) for feasible (w,b,ξ) of (6) and (α,b,ξ) of (10).

$$\begin{aligned} &\tfrac{1}{2}\alpha^T Q\alpha+\tfrac{v}{2}\xi^T\xi=\tfrac{1}{2}w^T w+\tfrac{v}{2}\xi^T\xi \\ &\geq\tfrac{1}{2}(w^*)^T w^*+\tfrac{v}{2}\xi^T\xi=\tfrac{1}{2}(\alpha^*)^T Q\alpha^*+\tfrac{v}{2}\xi^T\xi \end{aligned} \quad . \tag{11}$$

Thus, (w^*,b^*,ξ^*) is the optimal solution to (6).

Having Theorem 1, we can obtain the optimal solution of (6) by solving (10). According to Corollary 1, we know (10) is equivalent to the following smooth formation for the kernel case:

$$\min_{\alpha,\gamma}\ \tfrac{1}{2}\alpha^T Q\alpha+\tfrac{v}{2}\| P_\beta(e-Q\alpha-\gamma Y)\|^2 \quad . \tag{12}$$

4 NSSVM Implementation

To apply the Newton algorithm, first we have to prove the convexity of (5) and (13) and figure out the corresponding gradient vectors and Hessian matrixes, which can be directly obtained by figuring out the first and twice order derivatives to program (13). The compurgation is easy, and therefore the detailed process is omitted.

Defining a diagonal matrix $D=Diag(\gamma)$ for any vector $\gamma\in R^n$, and introducing several denotations $t=e-(DAw-De\gamma), v=\exp(-\beta t), Q=DA, Y=De$ and $M=(I+\beta Diag(v).Diag(P_\beta(t))$, then we get the formulas of the gradient vector and the Hessian matrix of NSSVM in the linear space.

$$\nabla F_\beta(x) = \frac{1}{C}\begin{bmatrix} w \\ 0 \end{bmatrix} + \begin{bmatrix} -Q^T \\ -Y^T \end{bmatrix} Diag(P_\beta(t)) Diag(e+v)^{-1}) \cdot e \ . \tag{13}$$

$$\nabla^2 F_\beta(x) = \frac{1}{C}\begin{bmatrix} I\,0 \\ 0\,0 \end{bmatrix} + 2\begin{bmatrix} -Q^T \\ -Y^T \end{bmatrix} Diag(e+v)^{-2} M[-Q\ -Y] \ . \tag{14}$$

To obtain the formulas of NSSVM in the kernel space, we define $t = e - (DK(A, A^T)\alpha - De\gamma)$ and $Q = DK(A, A^T)$, then we obtain the formulas of the gradient vector and the Hessian matrix, in which v, Y and M are the same with those in the linear space.

$$\nabla F_\beta(x) = \frac{1}{C}\begin{bmatrix} \alpha \\ 0 \end{bmatrix} + \begin{bmatrix} -Q^T \\ -Y^T \end{bmatrix} Diag(P_\beta(t)) Diag(e+v)^{-1}) \cdot e \ . \tag{15}$$

$$\nabla^2 F_\beta(x) = \frac{1}{C}\begin{bmatrix} Q\,0 \\ 0\,0 \end{bmatrix} + 2\begin{bmatrix} -Q^T \\ -Y^T \end{bmatrix} Diag(e+v)^{-2} M[-Q\ -Y] \ . \tag{16}$$

Using the expressions of $\nabla F_\beta((\alpha^T, \gamma)^T)$ and $\nabla^2 F_\beta((\alpha^T, \gamma)^T)$, we can easily deduce Theorem 2.

Theorem 2. Programs (5) and (12) are both convex programs.

Having Theorem 2, we can apply Newton algorithm using Armijo stepsize to figure out the optimal solution.

5 Numerical Experiments

We demonstrate now the effectiveness and speed of NSSVM on several real word datasets from the UCI machine learning respiratory.

All the experiments are carried out on a PC with P4 CPU, 3.06 GHz, 1GB Memory using MATALAB 7.01. When computing the gradient vector $\nabla F_\beta((\alpha^T, \gamma)^T)$ and the Hessian matrix $\nabla^2 F_\beta((\alpha^T, \gamma)^T)$, the limits of $Diag(P_\beta(t)) \cdot Diag(e+v)^{-1})$ and $Diag(e+v)^{-2} \cdot (I + \beta Diag(v) \cdot Diag(P_\beta(t)))$ are used.

First, we carry out the experiments in the liner space. Three moderated sized datasets are used, the Bupa Liver, the Ionosphere and the Pima Indians. To further testify the advantages of NSSVM, we compare the ten-fold training and testing accuracies of NSSVM with RLP, $SVM_{||1||}$, $SVM_{||2||}$ and FSV, in terms of the number of iterations, the ten-fold training accuracies and the ten-fold testing accuracies.

Table 1. Comparisons with Various Methods

Ten-fold Training Accuracies, % Ten-fold Testing Accuracies,%					
Data	***NSSVM***	***RLP***	$SVM_{\|1\|}$	$SVM_{\|2\|}$	***FSV***
Bupa	**71.60**	68.98	67.83	70.57	71.21
	70.19	64.34	64.03	69.86	69.81
Iono.	**94.94**	94.78	88.92	92.96	94.87
	89.18	86.04	86.10	89.17	86.76
Pima	**78.13**	76.48	75.52	77.92	77.91
	77.99	76.16	74.47	77.07	76.96

The results are illustrated in table 1, in which the bold type numbers indicate the best results. Obviously, NSSVM has the highest ten-fold training and testing accuracies. On the Ionosphere data, It has a training and a testing accuracy about 0.16% and 3.14% higher than RLP; It has a training and a testing accuracy about 6.02% and 3.08% higher than $SVM_{\|1\|}$; It has a training and a testing accuracy about 1.98% and 0.01% higher than $SVM_{\|2\|}$; It has a training and a testing accuracy about 0.07% and 2.42% higher than FSV.

In fact, SSVM has been proven to be more effective than RLP, $SVM_{\|1\|}$, $SVM_{\|2\|}$, SOR, FSV, SMO and SVM^{light}[12], so in the following, we will compare the performances of NSSVM with SSVM on Bupa Liver to further demonstrate its effectiveness using Gaussian radial basis kernel.

Table 2. Comparisons with SSVM

Methods		**Arm.**	**Iter.**	**Time**	**Tr. Cor.**	**Ts. Cor.**
SSVM	Lin.	1	4.1	0.01	70.11%	68.08%
		0	4.3	0.01	70.08%	67.83%
	Ker.	1	2	0.21	100%	60.36%
		0	2	0.20	100%	60.29%
NSSVM	Lin.	1	4.4	0.01	**70.53%**	**69.56%**
		0	4.3	0.01	**70.37%**	**68.43%**
	Ker.	1	3.3	0.73	100%	**61.12%**
		0	3.6	0.67	100%	**60.91%**

In the above table, the bold type numbers indicate higher accuracies of NSSVM than SSVM.

Apparently, NSSVM has higher accuracies over SSVM both in the linear and kernel case, which proves that removing the bias term $\gamma^2/2$ from SSVM model do increase the accuracy, while its the number of iterations and training time remains almost unchanged. Although Armijo stepsize can guarantee the global and quadratic convergence of Newton algorithm, it can cause only slight increases in accuracies and a little longer training time, so in practice, it can be turned off. Using kernel function, we can obtain much better results using less numbers of iterations.

6 Conclusions

This paper investigates a new smooth formulation SSVM, which is an unconstrained smooth formulation for the standard 2-norm error soft margin SVM, and proposes Newton-Armijo algorithm to commutate the optimal solution. By using the 2-norm of the Lagrangian multipliers vector to replace that of the weight of the hyperplane or using the connotative relation between the primal and dual program, NSSVM can be easily extended to the kernel space without doubt. Numerical results show that NSSVM has higher accuracies over existing methods. Future work includes finding other smooth penalty functions or searching new efficient algorithm to solve the unconstrained smooth model.

Acknowledgement

The authors give sincere thanks to the kind editors and anonymous reviewers for their valuable suggestions and comments, which have helped to improve the manuscript considerably.

References

1. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (2000)
2. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
3. Yu, H., Han, J., Chang, K.C.: PEBL: Positive-example based learning for Web page classification using SVM. In: Proc. 8th Int. Cong. Knowledge Discovery and Data Mining, Edmonton, Canada (2002)
4. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
5. Chen, X., Qi, L., Sun, D.: Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities. Mathematics of Computation 67, 519–540 (1998)
6. Qi, L., Sun, D.: Smoothing Functions and Smoothing Newton Method for complementarity and Variational Ineauqlity Problems. Journal of Optimization Theory and Applications 113(1), 121–147 (2002)
7. Lee, Y.-J., Mangasarian, O.L.: SSVM: A Smooth Support Vector Machine for Classification. Computational Optimization and Applications 22(1), 5–21 (2001)
8. Hsu, C.-W., Lin, C.-J.: A simple decomposition method for support vector machines. Machine Learning 46, 291–314 (2002)
9. Yuan, Y.-b., Yan, J., Xu, C.-x.: Polynomial smooth support vector machine (PSSVM). Chinese Journal of Computers 28(1), 9–71 (2005) (in Chinese)
10. Lee, Y.J., Hsieh, W.F., Huang, C.M.: SSVR: A smooth support vector machine for insensitive regression. IEEE Transactions on Knowledge and Data Engineering 17(5), 5–22 (2005)
11. Yan-Feng, F., De-Xian, Z., Hua-Can, H.: Smooth SVM Research: A polynomial- based Approach. In: Proceedings of the Sixth Conference on Information, Communications & Signal Processing, pp. 1–5 (2007)
12. Jin-zhi, X., Jin-lian, H., Hua-qiang, Y., Tian-ming, H., Guang-ming, L.: Research on a New Class of Functions for Smoothing Support Vector Machines. Acta Electronica Sinica 35(2), 366–370 (2007)

Convergence of GCM and Its Application to Face Recognition

Kai Li, Xinyong Chen, Nan Yang, and Xiuchen Ye

School of Mathematics and Computer, Hebei University, Baoding 071002, China
{likai_njtu,xychenhbu,nyanghbu,xiuchenye}@163.com

Abstract. We mainly generalize consistency method in semi-supervised learning by expanding kernel matrix,denoted by GCM(Generalized Consistency Method), and study its convergence. Aimed at GCM,we give the detailed proof for condition of convergence. Moreover,we further study the validity of some variants of GCM. Finally we conduct the experimental study on the parameters involved in GCM to face recognition. Meanwhile, the performance of GCM and its some variants are compared with that of support vector machine methods.

Keywords: Face recognition, Semi-supervised learning, Kernel matrix, Convergence.

1 Introduction

Face recognition has attracted many researchers in the area of pattern recognition and computer vision because of its immense application potential[1].So within the last several years,numerous methods have proposed for face recognition.Afterward,reaserchers start to use the machine learning method for face recognition and present some face recognition's algorithms based on machine learning technique.

It is well known that machine learning includes the supervised learning, the unsupervised learning, and the semi-supervised learning. Given a dataset(x_1,y_1), $(x_2,y_2),\ldots,$ (x_N,y_N),where $x_i=(x_{i1},x_{i2},\ldots,x_{iN})$ is the properties vector of the i-th sample, element x_{ij} is the j-th properties value of the i-th sample, and this value can be a discrete or continuous one. The primary differences among above the three kinds of learning methods are whether the system has prior knowledge or not and how much prior knowledge there is in the system. It is supervised learning or unsupervised learning when all values of $y_i(i=1,2,\ldots,N)$in N sample points are known or unknown. However, labeled instances in real world are often difficult, expensive, or time consuming to obtain, meanwhile unlabeled data may be relatively easy to collect, which is rarely used, it is called semi-supervised classification learning [2]. In other words, when the values of very few sample points in N sample points are known, then it is semi-supervised learning. The main goal of semi-supervised classification learning is that how to use few labeled data and a large amount of unlabeled data to enhance the

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 273–281, 2010.

generalization performance of system.Up to now,researcher have presented many semi-supervised learning algorithms [3] [4] [5].

In this paper we mainly study the consistency method,expand this method to more generic measure(denoted GCM) and give the detailed proof about the important condition of generalized consistency method's convergence. Moreover,we apply GCM algorithm to face recognition.

2 Generalized Consistency Method

Given a data set $X = \{(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}, \ldots, x_n\}$, $x_i \in R^m$ and $y_i \in \{1, 2, \ldots, c\}$, note that points for $x_{l+1}, \ldots, x_n$ don't have label in data set X and the goal of learning is to predict the labels of the unlabeled data $(x_{l+1}, \ldots, x_n)$.In general,l is far less than n ($l<<n$).Let F denote the set of matrices with non-negative entries. A matrix $F = [F_1^T, \ldots, F_n^T]$ corresponds to a classification on the dataset X by labeling each point x_i as a label. We can get the last label of each point by taking the maximum value from each row in matrix F, namely $y_i = arg\ max_j\{F_{ij}\}$. In addition, define matrix $Y = (Y_{ij})$,where $Y_{ij} = 1$ If $y_i = j$ else $Y_{ij} = 0$ and $W = (W_{ij})$, where if $i \neq j$ then $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2))$ else $W_{ij} = 0$.Aimed at this, Zhou et al.[3] presented the CM algorithm. In this paper,we generalize this algorithm and name it as GCM(Generalized Consistency Method). In the following,we briefly give GCM framework.

Step 1 Calculate the similarity between each two points in the data set to form the affinity matrix $W = (W_{ij})_{n\times n}$, where $W_{ij} = f(d(x_i, x_j))$,f is a function,and $d(x_i, x_j)$is a measure between a pair of points.

Step2 Construct a matrix $S = D^{-1/2}WD^{-1/2}$, where D is a diagonal matrix and its (i,i)-element is equal to the sum of the i-th row of W. "-1/2" denotes the reciprocal of square root of element in D matrix.

Step3 Iterate the following formula until convergence:

$$F(t+1) = \alpha SF(t) + (1-\alpha)Y, \tag{1}$$

where α is a parameter in (0,1). Step4 Let $F*$ denote the limit of the sequence $F(t)$. Make the label of each point according to the following way:

$$y_i = arg\ max_{j\leq c}\{F_{ij}^*\}. \tag{2}$$

3 Condition of Convergence for Generalized Consistency Method

In GCM, to obtain the resulting classification,it mainly uses iterated equation (1).So convergence for GCM is very important for its application.In ref. [3],Zhou et al. has shown that the absolute value of eigenvalues for matrix S is less than 1.However,they didn't give a detailed proof for condition of convergence of CM. In this section,we further study GCM and its variants. Especially,the detailed derivative procedures are given in this paper.

Lemma Given set $A = (a_{ij})_{n\times n}$ is a n-order complex matrix,its eigenvalue at least satisfies one of the following inequation [6]:

$$|\lambda - a_{ii}| \leq \sum_{1\leq j\leq n, j\neq i} |a_{ij}|, i = 1, 2, \ldots, n. \tag{3}$$

Proposition 1 The eigenvalues of matrix $S = D^{-1/2}WD^{-1/2}$ is in [-1, 1],where D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W.

Proof As $D^{-1}W = D^{-1/2}SD^{1/2}$, denote $D^{-1}W$ by P, so the matrix P is similar with S. As similarity matrix has the same eigenvalues, the eigenvalues of matrix S is equal to that of matrix P.Thus,we just need to prove that the eigenvalues of matrix P is in [-1,1]. As P is equal to $D^{-1}W$, we get the following expression:

$$P = \begin{pmatrix} P_{11} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & P_{nn} \end{pmatrix} \begin{pmatrix} 0 & W_{12} & \ldots & W_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ W_{n1} & W_{n2} & \ldots & 0 \end{pmatrix},$$

where $P_{ii} = (\sum_{j=1,j\neq i}^{n} W_{ij})^{-1}$.Let P_i denotes the sum of the i-th row elements in matrix P, we have the following expression

$$P_i = \sum_{j=1,j\neq i}^{n} (W_{ij} \diagup \sum_{j=1,j\neq i}^{n} W_{ij}) = 1.$$

It may be known by above lemma that the eigenvalues of matrix P is in [-1, 1].That is to say that the eigenvalues of matrix S is also in [-1, 1].

In the following,aimed at the exponent measure in GCM,we give an explanation for proposition 1.

For simplicity,we fix the dimension for matrix to 3 and specify function f in GCM as exponent function.Since $S = D^{-1/2}WD^{-1/2}$,matrix S become

$$\begin{pmatrix} 0 & \frac{1}{\sqrt{W_{12}+W_{13}}}W_{12}\frac{1}{\sqrt{W_{21}+W_{23}}} & \frac{1}{\sqrt{W_{12}+W_{13}}}W_{13}\frac{1}{\sqrt{W_{31}+W_{32}}} \\ \frac{1}{\sqrt{W_{21}+W_{23}}}W_{21}\frac{1}{\sqrt{W_{12}+W_{13}}} & 0 & \frac{1}{\sqrt{W_{21}+W_{23}}}W_{23}\frac{1}{\sqrt{W_{31}+W_{32}}} \\ \frac{1}{\sqrt{W_{31}+W_{32}}}W_{31}\frac{1}{\sqrt{W_{12}+W_{13}}} & \frac{1}{\sqrt{W_{31}+W_{32}}}W_{32}\frac{1}{\sqrt{W_{21}+W_{23}}} & 0 \end{pmatrix}$$

Now let $T_i(i{=}1,2,3)$ denote sum of column i elements.Then

$$T_1 = \frac{1}{\sqrt{W_{12}+W_{13}}}W_{12}\frac{1}{\sqrt{W_{21}+W_{23}}} + \frac{1}{\sqrt{W_{31}+W_{32}}}W_{31}\frac{1}{\sqrt{W_{12}+W_{13}}}$$

As W is a symmetric matrix and $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$,we have $W_{ij} = W_{ji}$.On the basis of $W_{23}W_{13} \geq W_{21}$ and $W_{13} + W_{23} \geq 2\sqrt{W_{21}}$,we further obtain $T_1 \leq \frac{1}{2}(W_{21}^{1/4} + W_{31}^{1/4})$.Similar,we can obtain $T_2 \leq \frac{1}{2}(W_{12}^{1/4} + W_{32}^{1/4})$ and $T_3 \leq \frac{1}{2}(W_{13}^{1/4} + W_{23}^{1/4})$.By feature of exponent function,$T_i(i{=}1,2,3)$is all smaller than 1. According to lemma,we can know that proposition 1 holds true.

Proposition 2 Let F^* be the limit of the sequence $\{F(t)\}$, we have $F^* = (I - \alpha S)^{-1}Y$ for the classification [3].

It is noted that since P is a stochastic matrix of S and $P = D^{-1}W = D^{-1/2}SD^{1/2}$, $F^* = (I - \alpha S)^{-1}Y$ can be approximately expressed as $F^* = (I - \alpha P)^{-1}Y$ and $F^* = (I - \alpha P^T)^{-1}Y$ [3].The interested reader is referred to [3] about consistency method. In GCM,there exist similar conclusions.

4 Experiments

In this section, our experimental focus is to study effectiveness of the GCM algorithm and its variants through face database.Moreover, we experiment the parameter involved in GCM algorithm. Thus, we use Orl [7] and Yale face database[8]. Orl database contains 10 different images of each of 40 distinct subjects; the resolution of each image is 112*92 pixels. In all the images of Orl, the subjects are in an upright, frontal position with some variation in lighting, facial expression, facial details etc. And the Yale database contains 11 different images of each of 15 distinct subjects, the resolution of each image is 100*100 pixels.In all the images of Yale, the subjects have different occlusion and the lighting change significantly. A part of face images are shown in Fig.1 and Fig.2.

Fig. 1. A part of Orl face database

Fig. 2. A part of Yale face database

Aimed at scale factor σ of exponential function, we select 5 values which are 0.15,0.35,0.55,0.75 and 0.95(in experimental results,using $S_i(i$=1,2,. . . ,5)denote them), respectively. In experiments, each chosen number of samples with class label is the integer value between [1,10] and [1,11] respectively.The results reported here are the mean of 10 test runs under same scale factor. In addition, in order to compare different algorithm's performance, we select LIBSVM to experiment on the face database. For Orl, the training number chosen is 40, 80, 120,160, 200, 240,280,320 and 360, respectively and the remaining is used for testing.While for Yale, the training number chosen is 15, 30, 45, 60,75, 90, 105,120,135 and 150, respectively and the remaining is used for testing.In SVM, the selected kernel is radial basis function, other parameters use default values in LIBSVM. We run 10 times for SVM classifier. The results reported here is the mean of 10 test runs. The experimental results are shown in the Fig.3-Fig.8,where Fig.3-Fig.5 are the experimental results of Orl face database and

Fig.6-Fig.8 are the experimental results of Yale face database. In experimental results,Fig.3 and Fig.6 are the results of the standard GCM algorithm (denoted as GCMS algorithm),respectively.And Fig.4 and Fig.7, Fig.5 and Fig.8 are the results of a variant of GCM algorithm (denoted by GCMP and GCMPT algorithm),respectively.In other words, they are the results of GCM algorithm which replace S with P and P^T. From Fig.3-Fig.8 , we can see that there has not significantly change when replace S matrix with the P^T, while when replace S matrix with P, the test accuracy has a large fluctuations. Especially when $S9 = 0.95$, there have a significantly change. In addition, for different values about scale factor, GCM algorithm's accuracy are quite different; especially the value of scale factor is more sensitive under the small number of known class label samples.

To demonstrate the performance of GCM algorithm, we also use the SVM classifier for the face recognition. In experiments, we choose different number of face samples for training and the remaining for testing. The results are shown in Table 1 and Table 2, where Table 1 is the result of Orl database and Table 2 is the result of Yale database.

Table 1. Experimental results with different methods for Orl

Method	40	80	120	160	200	240	280	320	360
GCMS	0.6898	0.8022	0.8540	0.8837	0.9077	0.9255	0.9351	0.9443	0.9522
GCMP	0.5632	0.6796	0.7418	0.7770	0.8037	0.8186	0.8295	0.8403	0.8499
GCMPT	0.7221	0.8177	0.8712	0.9016	0.9194	0.9347	0.9441	0.9543	0.9583
SVM	0.1608	0.2656	0.3529	0.4079	0.4705	0.5238	0.5650	0.6075	0.66

Table 2. Experimental results with different methods for Yale

Method	15	30	45	60	75	90	105	120	135	150
GCMS	0.5479	0.6747	0.7374	0.7601	0.7834	0.7982	0.8148	0.0.8260	0.8363	0.8411
GCMP	0.4494	0.5739	0.6353	0.6698	0.6997	0.7142	0.7235	0.7331	0.7366	0.7513
GCMPT	0.5730	0.6741	0.7238	0.7545	0.7811	0.8	0.8173	0.8292	0.8425	0.8524
SVM	0.1960	0.2519	0.2992	0.3352	0.3678	0.352	0.345	0.3178	0.3533	0.28

From the experimental results, we can know that performance for GCM algorithm outperforms that of SVM algorithm, the performance of GCMS is equivalent to the that of GCMPT method and the results show a large deviation between GCMS and GCMP method. We think that the reason of the results showing a large deviation between GCMS and GCMP algorithm is that P matrix doesn't a symmetric matrix.Besides,for Yale face database,we experiment with Euclidean function about f for GCM,denoted GCMO.Table 3 shows experimental result.It is seen that performance of GCMO is very inferior to those of methods presented above.Now,we further study them.

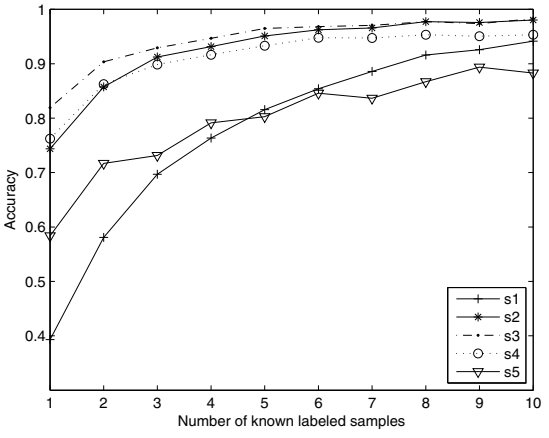


Fig. 3. Experimental result for GCMS in Orl, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

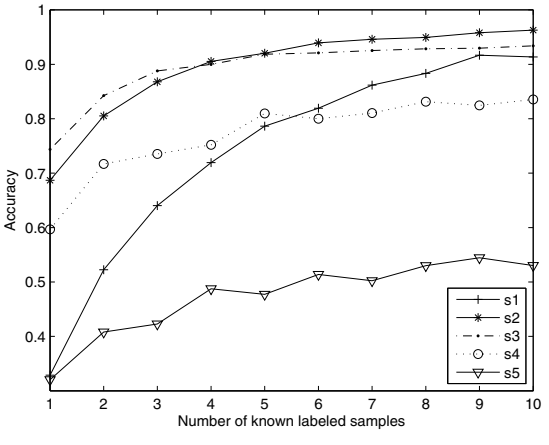


Fig. 4. Experimental result for GCMP in Orl, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

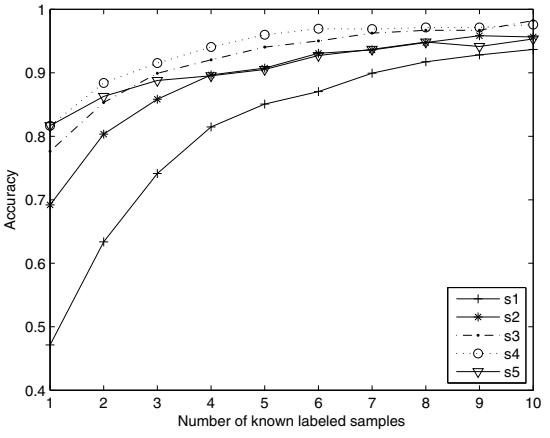


Fig. 5. Experimental result for GCMPT in Orl, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

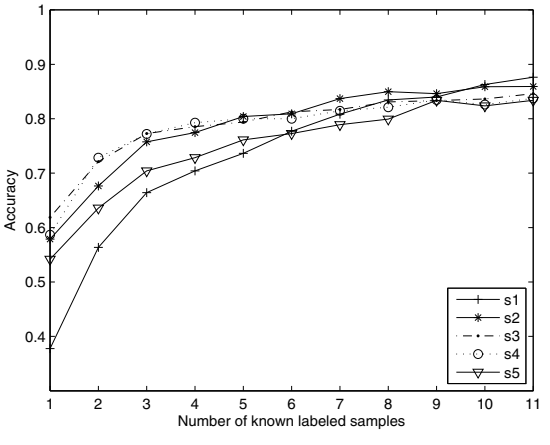


Fig. 6. Experimental result for GCMS in Yale, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

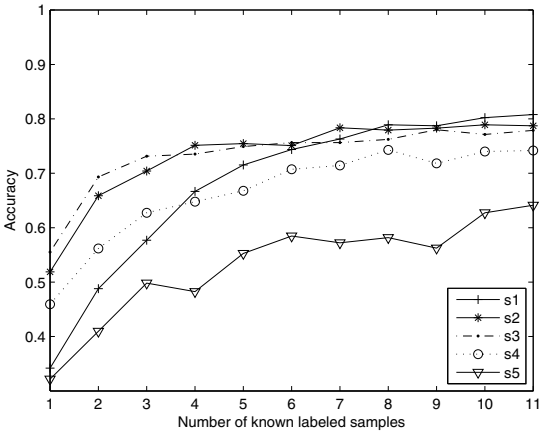


Fig. 7. Experimental result for GCMP in Yale, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

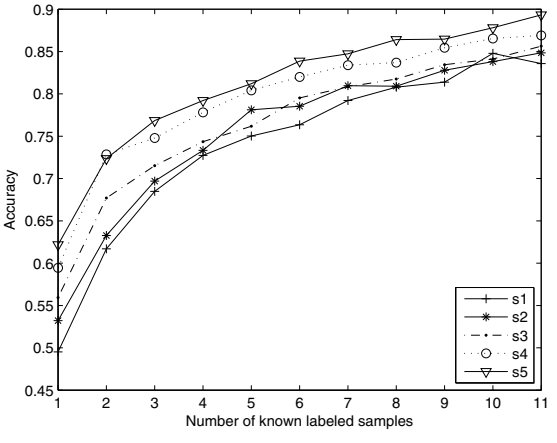


Fig. 8. Experimental result for GCMPT in Yale, where horizontal ordinate is the number of known labeled samples and vertical coordinate is tested accuracy.Function f for GCM is specified as exponent measure.

Table 3. Experimental results with GCMO for Yale

Method	15	30	45	60	75	90	105	120	135	150
GCMO	0.1453	0.2255	0.2934	0.3341	0.3249	0.2976	0.2969	0.2937	0.3009	0.3126

5 Conclusions

Generalized consistency method and its variants are deeply analyzed, the proof about the important condition for generalized consistency method's convergence is given in detail. In addition, this paper carries out deep research on the convergence of some other transformations of this algorithm, and proved that using the transformation matrix in GCM algorithm is reasonable. Finally we carry out the study on the parameters and the transformation matrix involved in this algorithm for face recognition. Meanwhile, performance of consistency method and its variants are compared with that of support vector machine.

Acknowledgments. The authors would like to thank Hebei Natural Science Foundation for its financial support (No: F2009000236).

References

1. Jiang, X.D., Mandal, B., Kot, A.: Eigenfeature Regularization and Extraction in Face Recognition. Pattern Recognition 30, 383–394 (2008)
2. Zhu, X.J.: Semi-supervised Learning Literature Survey. Computer Sciences TR 1530, University of Wisconsin - Madison (2008)
3. Zhou, D.Y., Bousquet, O., Thomas, N.L., et al.: Learning with Local and Global Consistency. In: Thrun, S., Saul, L., Scholkopf, B. (eds.) Advances in Neural Information Processing Systems 16, pp. 321–328. MIT Press, Cambridge (2004)
4. Yin, X.S., Chen, S.C., Hu, E.L., Zhang, D.Q.: Semi-supervised Clustering with Metric Learning: an Adaptive Kernel Method. Pattern Recognition 43, 1320–1333 (2010)
5. Mahdieh, S.B., Saeed, B.S.: Kernel-based Metric Learning for Semi-supervised Clustering. Neurocomputing 73, 1352–1361 (2010)
6. Su, Y.C., Jiang, C.B., Zhang, Y.H.: Theory of Matrix. Science Press, Beijing (2007)
7. Samaria, F., Harter, A.: Parameterization of a Stochastic Model for Human Face Identification. In: Proc. 2nd IEEE Workshop Applications of Computer Vision, Sarasota FL, pp. 138–142 (1994)
8. Yale University face database, http://cvc.yale.edu/projects/yalefaces/yalefaces.html

Designing a Multi-label Kernel Machine with Two-Objective Optimization

Hua Xu and Jianhua Xu

School of Computer Sceince and Technology
Nanjing Normal University
Nanjing, Jiangsu 210097, China
xuhuayg@163.com, xujianhua@njnu.edu.cn

Abstract. In multi-label classification problems, some samples belong to multiple classes simultaneously and thus the classes are not mutually exclusive. How to characterize this kind of correlations between labels has been a key issue for designing a new multi-label classification approach. In this paper, we define two objective functions, i.e., the number of relevant and irrelevant label pairs which are ranked incorrectly, and the model regularization term, which depict the correlations between labels and the model complexity respectively. Then a new kernel machine for multi-label classification is constructed using two-objective minimization and solved by fast and elitist multi-objective genetic algorithm, i.e., NSGA-II. Experiments on the benchmark data set Yeast illustrate that our multi-label method is a competitive candidate for multi-label classification, compared with several state-of-the-art methods.

Keywords: multi-label classification, kernel, regularization, multi-objective optimization, NSGA-II.

1 Introduction

Multi-label classification is a particular learning task, where some samples are associated with multiple classes at the same time, and thus the classes are not mutually exclusive. Currently there mainly exist two types of discriminative methods for multi-label classification: problem transform and algorithm adaptation [1]. The former converts training data sets to fit a large number of existing classification techniques, while the latter extends some existing multi-class algorithms to satisfy multi-label training data sets.

Problem transform methods split a multi-label training set into either one or more single label (binary or multi-class) subsets, train a sub-classifier for each subset using some popular approaches, and integrate all sub-classifier into an entire multi-label classifier. Now there are three widely used data decomposition strategies: label powerset (LP), one-versus-rest (OVR), and one-versus-one (OVO). LP strategy considers each possible label combination of more than one class in a multi-label data set as a new single class, and then builds a standard multi-class training data set. Its main disadvantage is that many classes have very few samples [1, 2]. OVR strategy divides a k-label training set into k binary subsets, in which the positive and negative

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 282–291, 2010.

samples in the *i*th subset come from the *i*th class and all other classes respectively. This strategy has been validated using many popular binary methods, e.g., support vector machines (SVM) [1-3], C4.5 [1, 4], naïve Bayes [1, 5], and *k*NN [1, 4]. OVO strategy splits a *k*-label set into $k(k\text{-}1)/2$ subsets, which results into some special subsets including a small proportion of double label samples. It seems reasonable to locate these double label samples between positive and negative samples. Two parallel hyper-planes are trained to separate three classes in [6], whereas in [7] these double label samples are forced to reside at the marginal region of binary SVM. The main limitation of problem transform methods is that they do not directly take the correlations between labels into account.

In contrast to problem transform methods, algorithm adaptation methods extend some specific multi-class approaches to deal with entire multi-label training sets directly, usually resulting into some optimization problems with very high computational complexity. Their main advantage is that the correlations between labels are considered explicitly. So far, many famous multi-class classifiers have been generalized to build their corresponding multi-label versions.

Through modifying the formula of entropy calculation and permitting multiple labels at the leave of the tree, a C4.5-type multi-label algorithm was proposed in [8]. BoosTexter [9] was derived from the well-known Adaboost algorithm, which include two slightly different versions: AdaBoost.MH and Adaboost.MR. The former is to predict the relevant set of labels of a sample, whereas the latter to rank labels of a sample in descending order. ADTboost [10] combines alternative tree with Adaboost.MH. It is hard to control the complexity of such boosting or tree based multi-label methods [3].

Nearest neighbor (*k*NN) or instance-based (IB) method has been extended to build two slightly different algorithms: ML-*k*NN [11, 12] and IBLR-ML [4], cascading standard *k*NN and Bayesian inference. They utilize different ways to estimate the posterior probability for each label, in which ML-*k*NN uses Bayesian rule and IBLR-ML logistic regression.

In principle, traditional back-propagation (BP) neural networks can deal with multi-label classification directly via assigning many ones at their output layer. To improve the performance of BP, its multi-label version (BP-MLL) [13] utilizes a new empirical loss function derived from the ranking loss and a regularization term. Rank-SVM [3] is a typical multi-label support vector machine, whose empirical loss function also is based on the ranking loss. It is noted that for BP-MLL and Rank-SVM, their objective functions combine an empirical loss function with a regularization term through weighted sum method. This is a classical solution method for multi-objective optimization [14, 15].

In this paper, we regard multi-label classification as a multi-objective optimization problem rather than a single-objective one. On the basis of kernel based decision functions, two objective functions: the number of relevant and irrelevant label pairs which are ranked incorrectly, and the model regularization term, are minimized at the same time using fast and elitist multi-objective genetic algorithm, i.e., NSGA-II [16, 17]. This setting can avoid determining a regularization constant in advance. Experimental results demonstrate that our new method is a highly competitive candidate for multi-label classification, compared with many state-of-the-art methods.

The remainder of this paper is organized as follows. In section 2, Rank-SVM is reviewed. Our new kernel machine for multi-label classification is proposed in section 3 and NSGA-II is summarized in section 4. Our algorithm is experimentally evaluated and compared with some other popular multi-label classification algorithms on the benchmark data set Yeast in section 5. Finally some conclusions are drawn.

2 Multi-label Support Vector Machine

In this section, we briefly review the famous multi-label support vector machine, i.e., Rank-SVM [3]. Assume a q-label training data set of size l to be

$$S = \{(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_i, Y_i), ..., (\boldsymbol{x}_l, Y_l)\}, \tag{1}$$

where $\boldsymbol{x}_i \in R^d$ and $Y_i \in 2^Y$ denote the ith training sample in the d dimensional real space and its relevant set of labels, and $Y = \{1, ..., q\}$ is a finite set of class labels. Additionally, we refer to the complement of Y_i, i.e., $\overline{Y}_i = Y \setminus Y_i$, as the irrelevant set of labels of $\boldsymbol{x}_i$. We define a linear decision function for each class, i.e.,

$$f_m(\boldsymbol{x}) = <\boldsymbol{w}_m, \boldsymbol{x}> + b_m, m = 1, ..., q, \tag{2}$$

where $\boldsymbol{w}_m$ and b_m stand for the weight vector and bias of the mth decision function. It is desired that the decision function values of relevant labels should be greater than those of irrelevant labels for every sample. Usually this objective is achieved by minimizing the ranking loss, which expresses the number of label pairs that the irrelevant labels are ranked higher than the relevant labels,

$$\text{RLoss} = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{|Y_i||\overline{Y}_i|}\left|(m,n) \in (Y_i \times \overline{Y}_i), \text{s.t.} f_m(\boldsymbol{x}_i) \le f_n(\boldsymbol{x}_i)\right|\right). \tag{3}$$

Rank-SVM builds a primary optimization problem as follows,

$$\begin{aligned} &\min \sum_{m=1}^{q} <\boldsymbol{w}_m, \boldsymbol{w}_m> + C\sum_{i=1}^{l}\frac{1}{|Y_i||\overline{Y}_i|}\sum_{(m,n)\in(Y_i\times\overline{Y}_i)}\xi_{imn}, \\ &\text{s.t.} \quad f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i) \ge 1 - \xi_{imn}, (m,n) \in (Y_i \times \overline{Y}_i), \\ &\qquad \xi_{imn} \ge 0, i = 1, ..., l, \end{aligned} \tag{4}$$

where C is a predetermined regularization constant which can control the tradeoff between the number of errors and the complexity of classifier. For a pair of relevant and irrelevant labels $(m \in Y_i, n \in \overline{Y}_i)$ of the ith sample, if $f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i) \ge 1$, the corresponding slack variable $\xi_{imn} = 0$; otherwise $\xi_{imn} > 0$. Therefore the second term of objective function in (4) is a proxy of the ranking loss essentially. Similarly to SVM and other kernel machines, the above problem (4) can be converted into its dual version and then inner product between two sample vectors can be replaced by various kernels satisfying the Mercer theorem [18].

3 Multi-label Kernel Machine with Two-Objective Optimization

We propose our multi-label kernel algorithm based on two-objective optimization in this section. It is observed that the objective function of (4) consists of two terms, the regularization term to reflect the complexity of classifier and the empirical loss function to characterize the ranking loss, and further two of them are combined using weighted sum method. Now we split such an objective function into two objective ones and construct a two-objective optimization problem for multi-label classification,

$$\begin{aligned} &\min \ \sum_{m=1}^{q} < \boldsymbol{w}_m, \boldsymbol{w}_m >, \\ &\min \ \sum_{i=1}^{l} \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(m,n)\in(Y_i\times\bar{Y}_i)} \xi_{imn}, \\ &\text{s.t.} \quad f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i) \geq 1 - \xi_{imn}, (m,n) \in (Y_i \times \bar{Y}_i), \\ &\qquad \xi_{imn} \geq 0, i = 1, ..., l. \end{aligned} \tag{5}$$

Conversely, we also regard (3) as a classical solution for (5) through weighted sum method [14, 15]. It is attractive that the constant C is not needed in (5). Note that in (5), two objectives conflict with each other, which means that we can not achieve their minima at the same time. Since there are a large number of variables to be solved, it is difficult to deal with (5) directly using existing multi-objective optimization techniques. We rewrite the constraints in (5) as

$$\xi_{imn} = \begin{cases} 0, \ if \ f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i) \geq 1, \\ 1 - f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i), \ \text{otherwise.} \end{cases} \tag{6}$$

Accordingly we utilize the definition of ranking loss (3) to replace the second objective function in (5),

$$\min \ \sum_{i=1}^{l} \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(m,n)\in(Y_i\times\bar{Y}_i)} \theta(\xi_{imn}), \tag{7}$$

where $\theta(z)$ stands for the step function, i.e., $\theta(z) = 1$ if $z > 0$; otherwise $\theta(z) = 0$. Now, a two-objective optimization problem for multi-label classification can be constructed as,

$$\begin{aligned} &\min \ \sum_{m=1}^{q} < \boldsymbol{w}_m, \boldsymbol{w}_m >, \\ &\min \ \sum_{i=1}^{l} \sum_{(m,n)\in(Y_i\times\bar{Y}_i)} \theta(1 - f_m(\boldsymbol{x}_i) - f_n(\boldsymbol{x}_i)). \end{aligned} \tag{8}$$

Here it is worth noting that the coefficient in the front of the second sum in (7) has been neglected to derive a concise formula. In this case, the second objective function in (8) can be regarded as the number of relevant and irrelevant label pairs, which are

ranked incorrectly. To design a nonlinear kernel multi-label classification algorithm, q kernel-based decision functions are adopted,

$$f_m^k(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_{im} k(\boldsymbol{x}_i, \boldsymbol{x}) + b_m, m = 1, ..., q, \tag{9}$$

where $k(\boldsymbol{x}_i, \boldsymbol{x})$ denotes some kernel function [18], and $\boldsymbol{\alpha}_m = [\alpha_{1m}, ..., \alpha_{lm}]^T$ and b_m stand for the coefficients vector and bias of the mth kernel decision function. In this paper, the polynomial and RBF kernels are utilized,

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{y}) &= (\boldsymbol{x}^T \boldsymbol{y})^p, \\ k(\boldsymbol{x}, \boldsymbol{y}) &= \exp(-\gamma \left\| \boldsymbol{x} - \boldsymbol{y} \right\|^2), \end{aligned} \tag{10}$$

where p and γ denote the degree of polynomial kernel and the scale factor of RBF kernel respectively, and $\left\| \boldsymbol{x} - \boldsymbol{y} \right\|$ stands for the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. Similarly to the linear form (8), a two-objective optimization problem for our multi-label kernel algorithm is constructed as follows,

$$\begin{aligned} &\min \sum_{m=1}^{q} < \boldsymbol{\alpha}_m, \boldsymbol{\alpha}_m >, \\ &\min \sum_{i=1}^{l} \sum_{(m,n) \in (Y_i \times \overline{Y}_i)} \theta(1 - f_m^k(\boldsymbol{x}_i) - f_n^k(\boldsymbol{x}_i)). \end{aligned} \tag{11}$$

We will solve (11) using NSGA-II outlined in the next section, producing a Pareto optimal set, each solution of which is impossible to improve in any objective without a simultaneous worsening in some other objectives. For (11), each solution consists of $(\boldsymbol{\alpha}_m, b_m), m = 1, ..., q$, whose size is $(l+1)q$. We pick up one solution from the Pareto set as our multi-label classifier, which minimizes the second objective function first and then minimizes the first objective one. In order to search a proper threshold to determine the relevant set of labels for a test sample, a linear model is built using least square regression method in the training procedure as in [3, 13]. In this paper, we simply refer to our novel multi-label kernel machine with two-objective optimization as ML-2OKM.

4 Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II

Nowadays, there exist many multi-objective evolutional algorithms which can solve our two-objective problem (11) [14, 15], among which fast and elitist multi-objective genetic algorithm, i.e., NSGA-II [17], is one of the most popular approaches, Its computational complexity is $O(MN^2)$, in which M is the number of objectives and N is the population size. In this section we briefly describe the main loop of NSGA-II.

A parent population P_0 of size N is initially created randomly. The population is sorted according to non-domination. Each solution is assigned a fitness or rank equal to its non-domination level in which 1 stands for the best level, 2 is the next best

level, and so on. In this case, minimization of fitness is assumed. To begin with, the binary tournament selection, recombination and mutation operators are utilized to create an offspring population Q_0 of size N.

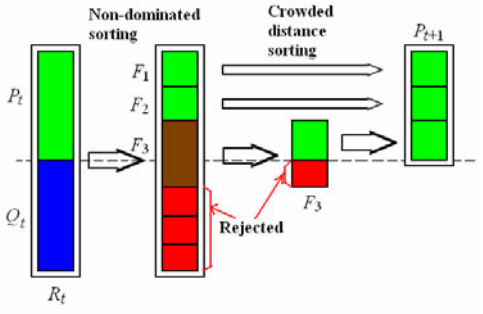


Fig. 1. Schematic of the NSGA-II. The combination of parent and offspring population ensures elitist selection, the non-dominated sorting ensures progress towards the Pareto-optimal front, and the crowded distance sorting ensures diversity among obtained solutions [16, 17].

From the first generation onward, the procedure becomes different. First, a combination population $R_t = P_t \cup Q_t$ is formed, whose size is $2N$. Then the population R_t is sorted according to non-domination. Solutions belonging to the best non-dominated set F_1 are of best solutions in R_t and must be emphasized more than any other solution in R_t. When the size of F_1 is smaller than N, its all members are chosen for the new population. The remaining members of the population are from subsequent non-dominated fronts in the order of their ranking. Thus solutions from the set F_2 are chosen next, followed by solutions from the set F_3, and so on. This procedure is continued until no more sets can be accommodated. Generally the count of solutions in the all sets from F_1 to F_l would be larger than the size of population. In order to pick up N members exactly, we sort the solutions in the last front F_l using the crowded comparison operator in descending order and choose the best solutions needed to fill all population members. Finally, the new population P_{t+1} of size N is used for selection, crossover and mutation to create a new population Q_{t+1} of size N.

Finally, the NSGA-II provides us a Pareto optimal set of solutions to implement our ML-2OKM. The NSGA-II procedure is shown in Fig.1. Please see the detailed description of NSGA-II in [17, 18]. It is worth noting that the NSGA-II uses float coding method rather than binary one, whose free C software is available at [19].

5 Experiments

In this section, we compare our method with several existing approaches on the benchmark data set Yeast. This data set comprises 2417 genes, in which 1500 and 917 genes are used for training and test. Each gene is originally described using micro-array expression data and phylogenetic profile, from which 103 features are extracted, and is associated with a subset of 14 functional classes. The 4.24 average labels indicate that this set is a typical multi-label classification problem. Before presenting

our experimental results, we list some evaluation measures used in this section in Table 1 briefly, in which the up (or down) arrow means that the higher (or lower) the corresponding measure is, the better a multi-label classification algorithm performs. For the detailed definitions of these measures, please refer to [1-3].

Table 1. Some evaluation measure

Measure	Abbreviation	Good performance
Ranking loss	RLoss	↓
One error	1Err	↓
Precision 1	Prec1	↑
Hamming loss	HLoss	↓
Accuracy	Acc	↑
Recall	Rec	↑
Precision 2	Prec2	↑

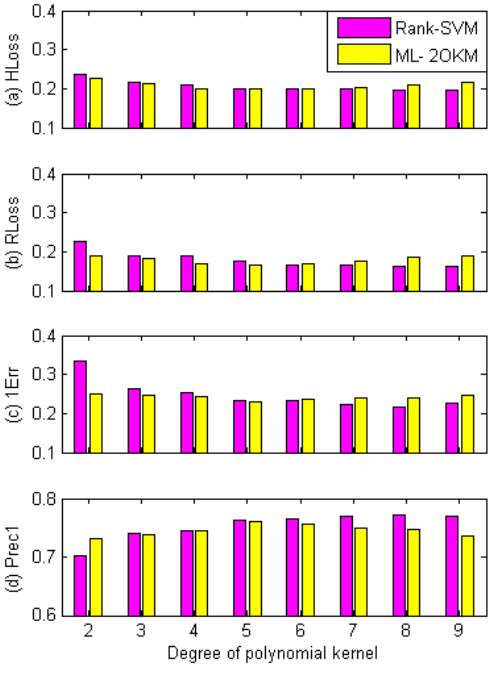


Fig. 2. A comparison between our ML-2OKM and Rank-SVM using the polynomial kernel

In [4, 5, 12, 13], 10-fold cross validation was used to evaluate the performance of several method on the Yeast. But, in this paper, we train our ML-2OKM using the training set only and then validate its performance on the test set. To the best of our knowledge, some experimental results based on the same experimental setting are recited directly from a large number of literatures.

Since our ML-2OKM is derived from Rank-SVM, we replicate the experimental setup of Rank-SVM in [3] to conduct the experimental evaluation of our method on

the Yeast. The degree of polynomial kernel varies from 2 to 9. It is worth noting that our method does not need a pre-defined regularization constant. For the NSGA-II, its key parameters include the number of generations and the size of populations, which are set as [100-200] and [80-150] respectively. Additionally, four same measures in [3] are estimated, as shown in Fig. 2. It is observed that our method performs better on the lower degree of polynomial kernel (2-5), whereas Rank-SVM works better on the higher degree of polynomial kernel (6-9). This means that our method builds simpler model than Rank-SVM does.

In [11], multi-label *k*NN (ML-*k*NN) was compared with three previous methods: Rank-SVM [3], ADTBoost [10] and BoostTexter [9]. IBLR-ML proposed in [4] was evaluated by using the test set in [20]. Here we directly list their results in Table 2, along with the best results of our ML-2OKM from Fig.2. Note that all methods in Table II are categorized into algorithm adaptation ones. In Table 2, our ML-2OKM obtains the lowest ranking loss. Although Rank-SVM achieves the best performance on the other three measures, the differences between Rank-SVM and our method are very small.

Table 2. A Comparison between our ML-2OKM and several algorithm adaptation methods

Method	Evaluation measue			
	HLoss↓	RLoss↓	1Err↓	Prec1↑
Rank-SVM[11]	**0.196**	0.179	**0.225**	**0.763**
ADTBoost[11]	0.213	--	0.245	0.738
BoosTexter[11]	0.237	0.298	0.302	0.717
ML-*k*NN[11]	0.197	0.168	0.239	0.761
IBLR-ML[20]	0.199	0.169	0.232	0.760
ML-2OKM(our)	0.198	**0.166**	0.230	0.760

Table 3. A Comparison between our ML-2OKM and some problem transform methods

Method	Evaluation measure			
	HLoss↓	Acc↑	Rec↑	Prec2↑
LP-kNN[1]	0.229	0.495	0.628	0.596
LP-C4.5[1]	0.286	0.399	0.528	0.529
LP-NB[1]	0.243	0.464	0.608	0.575
LP-SVM[1]	0.206	0.530	0.672	0.615
OVR-*k*NN[1]	0.243	0.479	0.601	0.596
OVR-C4.5[1]	0.259	0.423	0.593	0.561
OVR-NB[1]	0.301	0.421	0.531	0.610
OVR-SVM[1]	0.200	0.502	**0.711**	0.579
OVO-3CSVM[6]	0.199	0.537	0.654	0.693
OVO-DLSVM[7]	**0.181**	**0.572**	0.687	0.606
ML-2OKM(our)	0.202	0.511	0.577	**0.717**

We also compare our methods with three types of problem transform methods using the other four evaluation measures, as shown in Table 3. LP and OVR strategies are combined with *k* nearest neighbor (*k*NN), C4.5 decision tree, naïve Bayes (NB) and SVM. These eight multi-label algorithms were verified in [1]. In [6, 7], three class SVM (3CSVM) and double label SVM (DLSVM) were designed to deal with the double label binary data subsets from OVO decomposition strategy. In Table 3, all SVM-type multi-label classifiers are tested by the RBF kernel, in which our method

uses $\gamma = 2$. According to this table, OVO-DLSVM performs the best on Hamming loss and accuracy, and OVR-SVM and our ML-2OKM work the best on recall and precision 2 respectively.

According to the above experimental results, it is concluded that our ML-2OKM is a competitive multi-label classification algorithm, compared with many other state-of-the-art ones.

6 Conclusions

In this paper, we propose a novel multi-label kernel classification machine with two-objective optimization (simply ML-2OKM), which minimizes the number of relevant and irrelevant label pairs which are not ranked correctly and the model regularization term at the same time. Experimental results illustrate that our ML-2OKM can achieve a satisfactory performance, compared with some existing multi-label approaches. To the best of our knowledge, our method is the first multi-label kernel machine based on multi-objective optimization, which could open a new way to design new algorithms for multi-label classification. Additionally, our ML-2OKM only involves the kernel parameters, which will simplify model selection or parameter tuning task.

In this paper, we only validate our method using the famous data set Yeast. In our further work, we will test more benchmark data sets. Additionally fast and elitist multi-objective genetic algorithm (NSGA-II) can provide a Pareto optimal set (i.e., many classifiers), so we will create an ensemble of these classifiers for multi-label classification to improve the performance of our method further.

Acknowledgments. This work is supported by National Natural Science Foundation of China under grant no. 60875001. We also thank Deb K whose software tool NSGA-II is adopted to implement our algorithm.

References

1. Tsoumakas, G., Katakis, I.: Multi-label Classification: an Overview. International Journal of Data Warehousing and Mining 3(3), 1–13 (2007)
2. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning Multi-label Scene Classification. Pattern Recognition 37(9), 1757–1771 (2004)
3. Elisseeff, A., Weston, J.: A Kernel Method for Multi-labelled Classification. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems (NIPS 2002), vol. 14, pp. 681–687. MIT Press, Cambridge (2002)
4. Cheng, W.W., Hullermeier, E.: Combining Instance-based Learning and Logistic Regression for Multi-label Classification. Machine Learning 76(2/3), 211–225 (2009)
5. Zhang, M.L., Pena, J.M., Robles, V.: Feature Selection for Multi-label Naïve Bayes Classification. Information Science 179(19), 3218–3229 (2009)
6. Wan, S.P., Xu, J.H.: A Multi-label Classification Algorithm Based on Triple Class Support Vector Machine. In: The 2007 International Conference on Pattern Recognition and Wavelet Analysis (ICPRWA 2007), pp. 1447–1452. IEEE Press, New York (2007)

7. Li, J.Y., Xu, J.H.: A Fast Multi-Label Classification Algorithm Based on Double Label Support Vector Machine. In: The 2009 International Conference on Computational Intelligence and Security (CIS 2009), vol. 2, pp. 30–35. IEEE Press, New York (2009)
8. Clare, A., King, R.D.: Knowledge Discovery in Multi-label Phenotype Data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS(LNAI), vol. 2168, pp. 42–53. Springer, Heidelberg (2001)
9. Schapire, R.E., Singer, Y.: Boostexer: a Boosting-based System for Text Categorization. Machine Learning 39(1/2), 135–168 (2000)
10. Comite, F.D., Gilleron, R., Tommasi, M.: Learning Multi-label Alternative Decision Tree from Texts and Data. In: Perner, P., Rosenfeld, A. (eds.) MLDM 2003. LNCS, vol. 2734, pp. 35–49. Springer, Heidelberg (2003)
11. Zhang, M.L., Zhou, Z.H.: A k-nearest Neighbor Based Algorithm for Multi-label Classification. In: The 1st IEEE International Conference on Granular Computing (GrC 2005), pp. 718–721. IEEE Press, New York (2005)
12. Zhang, M.L., Zhou, Z.H.: ML-kNN: a Lazy Learning Approach to Multi-label Learning. Pattern Recognition 40(5), 2038–2048 (2007)
13. Zhang, M.L., Zhou, Z.H.: Multi-label Neural Networks with Application to Functional Genomics and Text Categorization. IEEE Transactions on Knowledge and Data Engineering 18(10), 1338–1351 (2006)
14. Konak, A., Coit, D.W., Smith, A.E.: Multi-objective Optimization Using Genetic Algorithms: a Tutorial. Reliability Engineering and System Safety 91(9), 992–1007 (2006)
15. Handl, J., Kell, D.B., Knowles, J.: Multiobjective Optimization in Bioinformatics and Computational Biology. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4(2), 279–292 (2007)
16. Srinivas, N., Deb, K.: Multiobjective Fucntion Optimization Using Nondominated Sorting Genetic Algorithms. Evolutional Computation 2(3), 221–248 (1995)
17. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
18. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
19. Free NSGA-II Software, http://www.lania.mx/~ccoello/EMOO/EMOOsoftware.html
20. Tahir, M.A., Kittler, J., Mikolajczyk, K., Yan, F.: Improving Multi-Label Classification Performance by Using Ensemble of Multi-label Classifiers. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 11–21. Springer, Heidelberg (2010)

Collision Detection Algorithm in Virtual Environment of Robot Workcell

Qian Ren[1], Dongmei Wu[2], Shuguo Wang[2], Yili Fu[2], and Hegao Cai[2]

[1] School of Information Science and Engineering, Harbin Institute of Technology (Weihai), 264209 Weihai, China
[2] State Key Laboratory of Robotics and System, Harbin Institute of Technology, 150001 Harbin, China
renqianhit@hitwh.edu.cn, wdm@hit.edu.cn, xzxx@hit.edu.cn, ylfms@hit.edu.cn, hgcai@hit.edu.cn

Abstract. Rapid and accurate collision detection is necessary to robot programming and simulation system based on virtual reality. A collision detection algorithm based on the data structure of the pre-built scene graph is proposed. It is the hierarchical accurate method. The architecture of collision detection system is established according to the hierarchy of scene graph for robot workcell, and implementation process of the collision detection algorithm is given. The collision detection system includes four collision detection managers. Groups filter manager, objects filter manager and faces & objects intersection manager are used in the broad phase of the collision detection, and they improve the speed of the algorithm by AABB intersection test and the layer by layer filtration. Polygons intersection manager is used in the narrow phase, and it ensures the accuracy of the algorithm. The result of the 3D graphics simulation experiment proves the effectiveness and feasibility of the way introduced.

Keywords: virtual environment, robot workcell, collision detection, scene graph.

1 Introduction

Collision detection has been researched in many fields such as robotic and computer graphics for a long time. Whether a contact or penetration has happened in virtual environment (VE) is the basic issue. Many articles have given amount of valuable and practicable study on the problem [1-8]. But in the later years, with the rising of virtual reality and distributed interactive simulation, many researches focus on collision detection.

Collision often happens between objects because of users' interaction or the movement of objects in VE. To keep the reality of VE, the collisions should be detected in real time, and the corresponding collision responses should be calculated, then rendering results can be updated. Otherwise, penetration may happen, which will destroy the reality of VE and users' immersion [9].

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 292–300, 2010.

Collision detection is one of the critical techniques for robot programming and simulation system based virtual reality, it affects the correctness, reality and security of application program. Collisions should be detected in VE of robot workcell at following aspects:

Robot programming [10]**.** When operators program virtual robots in VE, they should know the contact relation between virtual manipulator and operational objects, and judge if the manipulator has moved to the desired operational position.

Robot path planning and checking up [11]**.** In the path planning of virtual robot, if the moving manipulator of virtual robot collides with the obstacles, then system should send the order to stop the robot and re-plan the movement of manipulator. Collision detection can help the generation of collision-free paths for robot. Before the planned movement orders were send to the actual robot, it is necessary to check collisions between the moving robot and the obstacles in its workspace. It can help the operators to check the paths of the robot.

Layout of robot workcell [12]**.** By adding the collision detection capability in true 3D simulation scene with the sense of depth, we can visualize the layout of workcell before purchasing devices. Hence, we can better know the workspace of robot, and establish the layout of the robot workcell quickly. Thereby the rationality of the layout is ensured.

By analyzing the hierarchy of VE for robot workcell, the hierarchy of AABBs for robot workcell and the hierarchy of the scene graph for the collision detection are obtained, and the architecture of the collision detection system is composed of four managers for collision detection is established. Groups filter manager, objects filter manager and faces & objects intersection manager are used in the broad phase of the collision detection, and polygons intersection manager is used in the narrow phase. AABBs (Axis-Aligned Bounding Boxes) intersection test is applied in the broad phase. The computation time is reduced by using the layer-by-layer filtration of the first three managers. The rapid and accurate collision detection in the virtual interactive environment is realized.

2 Hierarchical Collision Detection System Based on the Data Structure of the Scene Graph

2.1 Hierarchy of the Scene Graph for Robot Workcell

The software system is bases on WTK^{TM} (WorldToolkit) functions library of American Sense8 [13]. 3D models of the robots and environment devices are created using RGMS (Robot Geometric Modeling System) is developed by author. Then the models with format of graphics files acceptable for WTK are saved. These geometries models files are called by WTK, and they as nodes are added into the scene graph. These nodes compose an interrelated and inverse tree structure according to definite connection. This tree structure is called scene graph [14].

The hierarchy of the scene graph for robot workcell is shown in Fig. 1. M environment devices and n robots locate at first level of the scene graph. All

equipments except robots in the robot workcell are environment devices, such as machines, feeders, parts and tools. Environment device is complex polyhedron is composed of multiple objects. These objects locate at second level of the scene graph. On the other hand, robot is a machine which is composed of the joints and the links. For example, the Puma560 robot is composed of six links which are joined by six joints is shown in Fig. 2. Thus, Puma560 robot can be divided into six sub-mechanisms, and each link has become a sub-mechanism. These links locate at second level of the scene graph.

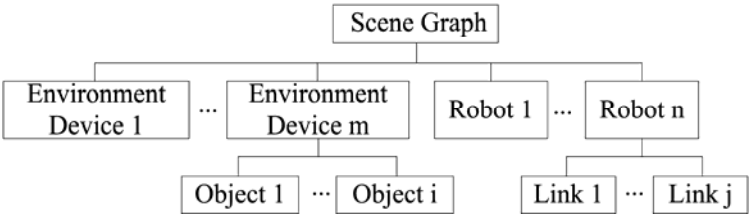


Fig. 1. Hierarchy of the scene graph for robot workcell

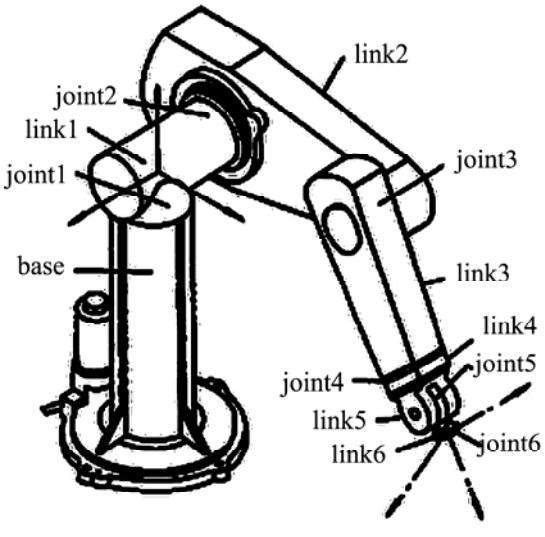


Fig. 2. Links and joints of robot PUMA560

2.2 Hierarchy of AABBs for Robot Workcell

According to the hierarchy of the scene graph for virtual robot workcell, hierarchical bounding boxes can be divided into two levels: the first level is entire AABBs, namely the AABBs of environment devices and the AABBs of robots; the second level is the AABBs of objects, namely the AABB of the objects which compose environment devices and the AABB of the links which compose the robot.

2.3 Hierarchy of the Scene Graph for Collision Detection

The hierarchy of the scene graph for collision detection is shown in Fig. 3. By establishing AABBs for robots, environment devices, objects and links, tree hierarchy of bounding boxes is formed, it can accelerate collision detection. Thus, the hierarchy of the scene graph for collision detection includes three levels: first level is groups which are AABBs of environment devices and AABBs of robots; second level is objects which are AABBs of the objects that compose environment devices and AABBs of the links that compose robot; third level is the polygons, because of objects and the links of robot can be expressed as the sets of polygons.

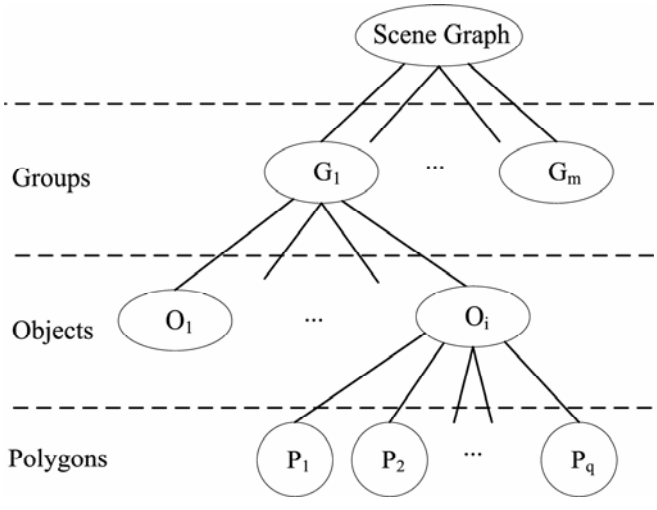


Fig. 3. Hierarchy of the scene graph for collision detection

2.4 Architecture of Collision Detection System

Collision detection manager is composed of four layers: groups filter manager (responsible for the intersection check between AABBs of environment devices and AABBs of robots), objects filter manager (responsible for the intersection check between AABBs of the objects and AABBs of the links), faces & objects intersection manager (responsible for the intersection check between AABBs of objects and polygons), polygons intersection manager (responsible for the intersection check between the polygons which compose the objects). Accordingly, the architecture of the collision detection system is shown in Fig. 4.

3 Implementation of the Collision Detection Algorithm

A virtual robot workcell is composed of multiple robots and much environment devices. A robot is a special movable mechanism with multiple joints. Although the

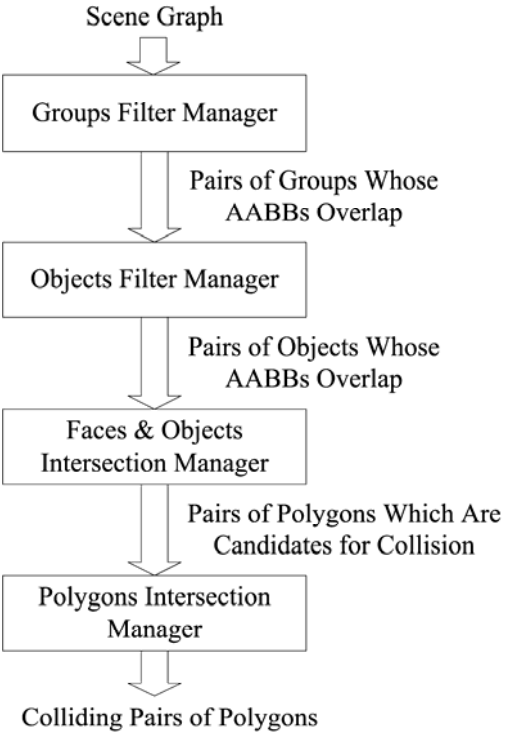


Fig. 4. Architecture of collision detection system

adjacent links of robot cannot collide each other due to the constraint at the moving range, it is possible to collide between the distant links, the end manipulator and the base. So, the implementation order for the collision detection algorithm is given here. Detailed implementation process of the collision detection algorithm is as follows.

3.1 Groups Filter Manager

This manager determines possible group candidates for collision, filtering groups that cannot collide. Two groups are called $G_1 = \{O_i \mid i = 1,...n\}$ & $G_2 = \{O_j \mid j = 1,...m\}$, if the corresponding AABBs do not overlap, then G_1 and G_2 cannot intersect (O_i and O_j are the object or link which composes the group G_1 and G_2 respectively), and this pair of groups is filtered out. Otherwise they became candidates for collision into objects filter manager. Such process is continued until all combinations are completed. After the collision detection for the AABBs of groups is completed, many objects which are separate each other and cannot intersect are filtered out, so the numbers of pairs of objects to need intersection calculation are reduced observably.

3.2 Objects Filter Manager

Before the collision detection for the pairs of objects, it is necessary to determine the type of the pairs of groups which output from groups filter manager and intersect. Following two cases exist here:

1) If they are two environment devices, then the overlap between the AABB of each object from one environment device and the AABB of each object from other environment device is checked. Such process is continued until all combinations are completed. This process filtered out pairs of objects whose AABBs do not overlap.

2) If one among two groups is a robot, because of that it is possible to a robot oneself collides, then the algorithm checks for the overlap between the AABBs of robot links, except using the method 1) to analyze the overlap between the combinations of the AABB of each object from two groups. Namely, the overlap between the combinations of AABBs of the robot links should be checked, but the overlap between the AABBs of adjacent links should not be checked.

Object filter manager determinates possible objects candidates for collision, filtering objects that cannot collide. Two objects called O_1 & O_2, if the corresponding AABBs do not overlap, then O_1 and O_2 cannot intersect, and this pair of objects is filtered out. Otherwise they became candidates for collision into faces & objects intersection manager.

3.3 Faces and Objects Intersection Manager

Candidates for collide into this manager are pairs of objects whose AABBs overlap. Assume that, such a pair of objects is $O_1 = \{P_i \,|\, i = 1,...n\}$ & $O_2 = \{P_j \,|\, j = 1,...m\}$, where, P_i & P_j are polygons that compose the object O_1 & O_2, respectively. The collision detection in this manager is comprised the two steps:

1) This manager checks for the intersection between all polygons P_i that belong to the object O_1 and the AABB of the object O_2, then $A_1 = \{S_k \,|\, k = 1,...p\}$ which is a set of polygons that belong to O_1 and intersect with AABB of O_2 is obtained, where, $S_k \in \{P_i \,|\, i = 1,...n\}$;

2) This manager checks for the intersection between all polygons P_j that belong to the object O_2 and the AABB of the object O_1, then $A_2 = \{S_l \,|\, l = 1,...q\}$ which is a set of polygons that belong to O_2 and intersect with AABB of O_1 is obtained, where, $S_l \in \{P_j \,|\, j = 1,...m\}$.

If the number of the set A_1 or A_2 is zero, then two objects called O_1 & O_2 don't intersect, otherwise they became candidates for collision into polygons intersection manager.

Collision detection in this manager makes ready for intersection check in polygons intersection manager. After collision detection is completed in this layer, calculation of collision detection in polygons intersection manager will be reduced observably.

3.4 Polygons Intersection Manager

This manager check for the polygons intersection between all polygons S_k that belong to the set of the polygons $A_1 = \{S_k \,|\, k = 1,...p\}$ and all polygons S_l that belong to the set of the polygons $A_2 = \{S_l \,|\, l = 1,...q\}$, then $A_3 = \{D_r \,|\, r = 1,...n\}$ which is the set of pairs of polygons that intersect each other is obtained, where, $D_r = (S_k, S_l)$ is a pair of polygons. If the number of the set A_3 is zero, then the polygons intersection does not happen between objects O_1 and O_2, otherwise the collision between two objects is determined, then the system will process the collision response.

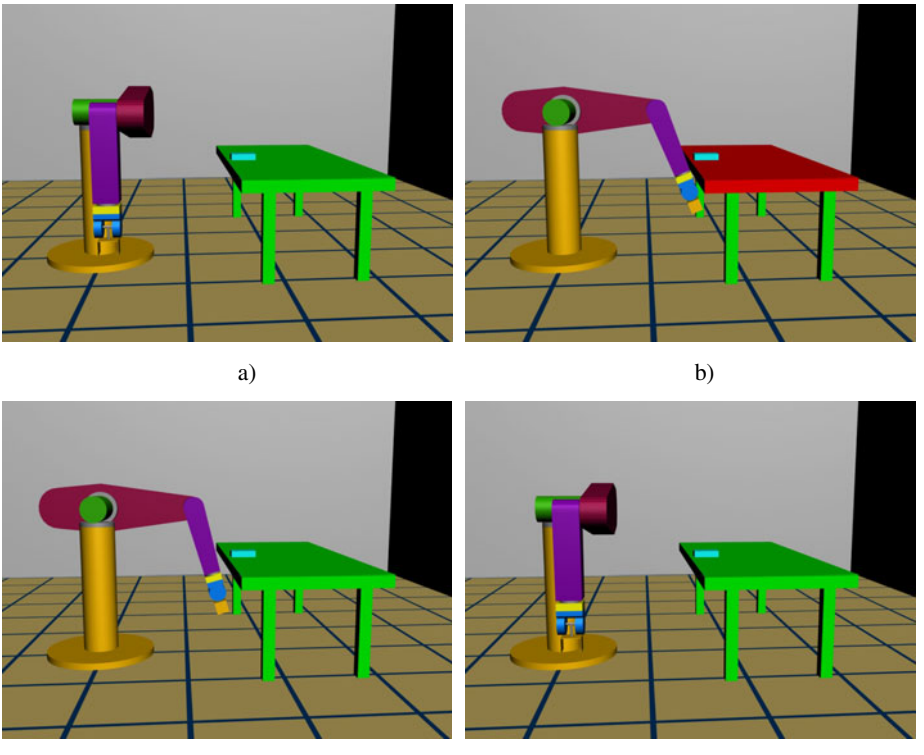


Fig. 5. Graphical simulation instance of the collision detection algorithms

4 Solution and Graphical Simulation Application Instance

We developed the prototype named RVTPS (Robot Virtual Teach Programming System) [15]. The hardware for this prototype includes a graphics workstation (SGI Indigo2™), a data glove (5th Glove™ of 5DT), a 3D tracker for hand position (3SPACE FASTRAK™ of Polhemus) and a LCD shutter glasses (CrystalEyes™

stereoscopic system of StereoGraphics). The software system is based on WTK™ (WorldToolkit) functions library of American Sense8.

Collision detection can be implemented with above proposed algorithm in robot off-line programming simulation. When a robot moves along a path, collision between the robot and environment devices can be detected as long as such tests are used at each discrete points of the path. Thus we can determine whether the current robot path is security. When the moving robot collides with environment devices, the simulation system will alarm: colliding parts are shown in red. Here, if the operator continues to send movement orders, then the simulator will refuse to enforce the order and alarm. Using this algorithm in the 3D graphics simulation, we can check whether robots collide with environment devices and movement logic between robots and other devices, and accordingly modify the robot program, until the robot's movement to meet the requirements.

The graphical simulation application instance of the algorithm is shown in Fig. 5. It expresses the process that collision between the robot and the environment devices is checked and processed by the system when we plan path of the robot. Fig (a) shows that the posture of the robot before the collision happens; Fig (b) shows: the system emits the collision alarm that the table top become red when the link of the robot collides with the table top; Fig (c) shows that the collision alarm disappear when the robot stop moving and begin return; Fig (d) shows that the robot return to the posture before the collision happens.

5 Conclusion

The collision detection technique in VE of robot workcell is researched. Hierarchical and step by step precise collision detection algorithm based the data structure of the scene graph is proposed in this paper. The implementation process of the algorithm in detail is described; finally the graphical simulation validates the effectiveness and feasibility of the algorithm. This algorithm is implemented on SGI graphics workstation, and it is applied in RVTPS developed by the author. The algorithm can successfully detect for the collisions between the elements in virtual robot workcell.

Acknowledgments. This work is supported by the Scientific Research Foundation of Harbin Institute of Technology (Weihai) under Grant HIT(WH).2007.21, and also supported by the National High-tech Research and Development Program (863 Program) under Grant 2006AA040103.

References

1. Kockara, S., Halic, T., Iqbal, K., Bayrak, C., Richard, R.: Collision Detection: A Survey. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 4046–4051. IEEE Press, New York (2007)
2. Martin, H., James, T.K., Joseph, S.B.M.: Evaluation of Collision Detection Methods for Virtual Reality Fly-throughs (1995), http://www.cosy.sbg.ac.at/~held/papers/cccg95_VR.ps.gz

3. Smith, A., Kitamura, Y., Takemura, H., Kishino, F.: A Simple and Efficient Method for Accurate Collision Detection Among Deformable Objects in Arbitrary Motion. In: Proceedings of the IEEE Virtual Reality Annual International Symposium, pp. 136–145. IEEE Press, Los Alamitos (1995)
4. Naylor, B.F.: Constructing Good Partitioning Trees. In: Proceedings of the Graphics Interface 1993, pp. 181–191. Canadian Information Processing Soc., Toronto (1993)
5. Garcia, A., Serrano, N., Flaquer, J.: Solving the Collision Detection Problem. IEEE Computer Graphics Application 14, 36–43 (1994)
6. Hubbard, P.M.: Approximating Polyhedra With Spheres for Time-critical Collision Detection. ACM Trans. Graphics 15, 179–210 (1996)
7. Barequet, G., Chazelle, B., Guibas, L.J.: BOXTREE: A Hierarchical Representation for Surfaces in 3D. Computer Graphics Forum 15, 387–484 (1996)
8. Klosowski, J.T., Held, M., Mitchell, J.S.B., Sowizral, H., Zikan, K.: Efficient Collision Detection Using Bounding Volume Hierarchies of K-DOPs. IEEE Transactions on Visualization and Computer Graphics 4, 1–22 (1998)
9. Jiaoying, S.: Virtual Reality Foundation and Practical Algorithms. Science Press, Beijing (2002)
10. Safaric, R., Sinjur, S., Zalik, B., Parkin, R.M.: Control of Robot Arm With Virtual Environment via the Internet. Proceedings of the IEEE 91, 422–429 (2003)
11. Jiacheng, T., Clapworthy, G.J.: Virtual Environments for Internet-based Robots. II. Path Planning. Proceedings of the IEEE 91, 389–395 (2003)
12. Jiaxin, W.: The Virtual Reality – Just not Only Simulation. The Robot Information 4, 3–7 (1993)
13. World Tool Kit Reference Manual. Sense8 Products Inc., Mill Valley, CA (1999)
14. Qian, R., Shuguo, W., Xiangli, C., Yili, F., Hegao, C.: Realistic Virtual Environment Generation for Robot Off-line Programming. Modular Machine Tool & Automatic Manufacturing Technique 3, 1–6 (2003)
15. Qian, R., Shuguo, W., Xiangli, C., Yili, F., Hegao, C.: Research on Robot Off-line Programming Based on Virtual Reality. ROBOT 25, 172–177 (2003)

A First Step towards Hybrid Visual Servoing Control Based on Image Moments

Xiaojing Shen[1], Dongmei Huang[1], and Xiaoxia Qin[2]

[1] Information Technology College, Shanghai Ocean University, HuCheng Circle Rd.999, 201306 Shanghai, China
[2] ZhongYe HuaTian Nanjing Automation Engineering co., ltd., AoTi Str. 69, 210019 Nanjing, China
{xjshen,dmhuang}@shou.edu.cn, qxxcdchz@163.com

Abstract. The paper is concerned with a specific class of visual servoing problem, in which camera motion (including both translation and rotation) are constraint to be in the Z-axis direction. Such a constraint condition makes it possible to find appropriate image moments reflecting object depth and orientation. Image moments, as a kind of global image features, can be benefit to the performance of visual servoing system, such as insensitivity to image noise, nonsingularity in image Jacobian, and etc. In the paper, the mathematic relationships between image moments and object-depth-and-orientation are firstly introduced. Then appropriate image moments are selected, on the basis of which a hybrid visual servoing system is build. In our system, visual servoing controller consists of two parts: one is called translation controller which is in charge of object depth control, the other is called rotation controller which controls object orientation. The simulation results show that our hybrid visual servoing system performances well with a high accuracy.

Keywords: visual servoing; image moment.

1 Introduction

In a visual servoing system, vision is employed as a feedback signal to control the actuator to track a static or moving part. Applications for visual servoing control include seam tracking, conveyor tracking, part placement precision, robotics, and etc. Generally speaking, the visual servoing control methods are divided into two classes [1]: position-based visual servo (PBVS) method and Image-based visual servo (IBVS) method. The former methods need to estimate the target position. The accuracy of this kind of methods relies on the accuracy of target model and camera calibration. While the latter methods treat the control problem as the one of controlling image features [2], hence can avoid the above disadvantage of PBVS methods. But IBVS methods have their own problems: local minima, coupled features that lead to unadequate robot trajectories [3], complexity control of a non-linear system, and etc.

To overcome the problems of IBVS and PBVS, some authors [4]-[5] have been recently studying the hybrid method of visual servoing. In most hybrid methods,

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 301–310, 2010.

translation motion v and rotation motion ω of a point in 3D workspace are separately controlled. The image Jacobian relationship is given by:

$$\dot{f} = J_v(u, v, Z)v + J_w(u, v)\omega$$

where $\dot{f}$ is the velocity of the corresponding point in the image, J_v (u, v, Z) is a function of both image coordinates of the point and its depth, J_w(u, v) has no relation to depth and is a function of only the image coordinates of the point.

It should be pointed out that the performances of visual servoing control largely depend on the selection of image features. Most hybrid visual servoing systems use local image feature, for example points (typically corners or centroids). The problems of local image feature are that they are sensitive to image noise, possibly result in singularities or poor conditioning in image Jacobian. One solution to such problems is using global image feature, since global image feature provide a generic representation of any object and is robust to image noise. Image moments [5], as a kind of global image feature, have been used for a long time in image analysis, and now can be calculated in real time. Studies indicate that object position and orientation in 3D space can be estimated from image moments [8].

In traditional IBVS control, motion in the Z-axis direction often play a key role, because depth information is essential to image Jacobian, and large rotation motion about Z-axis often leads to control problem. So, the study of our hybrid visual servoing control based on image moments is started from Z-axis control. That is, our study is concerned with a specific class of visual servoing problem, in which motion (involving translation and rotation) are constrained to be in the Z-axis direction. Therefore our control task can be described as how to design a hybrid controller under the above restrictions, in the control of which a camera is driven to its ideal position where the object depth and orientation are equal to the given expectation.

In the paper, firstly we try to find out proper image moments reflecting object depth and orientation. Then, based on the selected image moments we design a translation controller and a rotation controller, which lead to satisfying dynamics of a hybrid visual servoing system.

2 Mathematical Preliminaries

In this section, we analyze the mathematical relationship between image moments and object pose in the assumption of that:

1) the object be rigid and planar, which is always perpendicular to the optical axis of the camera (also called Z-axis or depth-axis) ;
2) in the process of visual servoing, only the Z-axis motion of the camera be permitted, which includes both translation component and rotation component.

As depicted in Fig.1, two reference frames are set: One is the camera frame {C} whose original point O is set at the center of camera lens, another is the image frame {I} whose original point o is set at the image center.

In Fig.1, an arbitary point P of the object is taken, and it projects at p in the image. Let P be at $[X,Y, Z]^T$ in the frame {C}, p be at $[x, y]^T$ in the frame {I}. In the assumption of the ideal pinhole model of camera, the perspective transformation formula is:

$$x = (\lambda k_x / Z)X \quad , \quad y = (\lambda k_y / Z)Y \tag{1}$$

where Z is in fact object depth, λ is camera focal length, kx and ky are respectively horizontal and vertical magniscales of the camera..

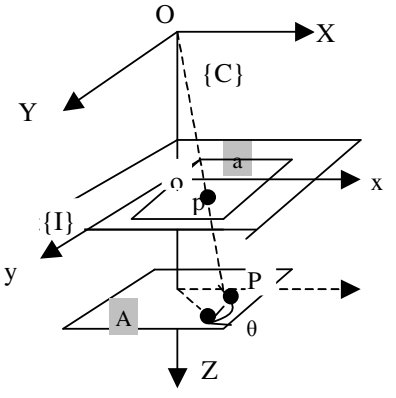


Fig. 1. The camera frame {C} and the image frame {I}

2.1 Image Moments

Let a be the region that an object projects on the image (the frame {I}), and a binary image function f(x,y) be the form of the following:

$$f(x, y) = \begin{cases} 1; (x, y) \in a \\ 0; (x, y) \notin a \end{cases}$$

Definition 1: the (p+q)th-order geometric moment of a binary image f(x,y) is defined by

$$M_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x, y)dxdy = \iint_{(x,y)\in a} x^p y^q dxdy \tag{2}$$

Definition 2: the (p+q)th-order central moment of a binary image f(x,y) is defined by

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q f(x, y)dxdy = \iint_{(x,y)\in a} (x-\bar{x})^p (y-\bar{y})^q dxdy \tag{3}$$

where $\bar{x}$ and $\bar{y}$ are given by the relations

$$\bar{x} = M_{10} / M_{00}, \quad \bar{y} = M_{01} / M_{00} \tag{4}$$

2.2 The Relationship between Zero-Order Image Geometric Moment and Object Depth

Supposing the object be static and the camera move translationally and rotationally in the Z-axis (the object motion in the frame {C} can be seen in fig.1). We have:

$$[X',Y',Z']^{T} = \mathbf{R}\,[X,Y,Z]^{T} + \mathbf{T} \tag{5}$$

where $[X',Y',Z']^T$ and $[X,Y,Z]^T$ represent the coordinates of a 3D point P in the frame {C} after and before camera motion, R is the rotation matrix and takes the form of

$$R = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

and **T is** the translation vector and expressed by:

$$\mathbf{T} = [\,0, 0, T_Z\,]^{T}, \quad T_Z = Z'-Z \tag{7}$$

If we observe in the frame {C}, we find θ is in fact the object orientation, and T_Z is the object displacement in the Z-axis.

Lemma 1: If an object is rigid and planar, and it is always perpendicular to the optic axis of a camera, the change of the zero-order image geometric moment ($\dot{M}_{00}$) has only related to the translational motion ($\dot{T}_Z$ or v_Z). And the relationship between the zero-order image geometric moment and the object depth can be described by:

$$M_{00}/M_{00}' = (Z'/Z)^2 \tag{8}$$

where M_{00}' and M_{00} denote the zero-order image geometric moments of the object at the depth Z' and Z.

Proof: Let $[X,Y,Z]^T$ and $[X',Y',Z']^T$ denote the coordinates of the point P in the frame{C} at time t and t', and $[x', y']^T$ and $[x, y]^T$ be the coordinates of the corresponding point in the frame {I} (seen in Fig.1). In the assumption of the section II, we can deduce the equation of (9) from (1), (5) and (6).

$$[x', y']^T = s\Lambda\Theta[X,Y]^T \quad , \quad [x, y]^T = \Lambda[X,Y]^T \tag{9}$$

where

$$s = \frac{Z}{Z'} \ , \ \Lambda = diag(\frac{\lambda f_x}{Z}, \frac{\lambda f_y}{Z}) \ , \ \Theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{10}$$

For the 2D region of a in the frame {I} (seen in Fig.1), we get the following equations of (11) from (2):

$$M_{00} = \iint_{(x,y)\in a} dxdy = \iint_{(X,Y)\in A} \left|\partial(x,y)/\partial(X,Y)\right| dXdY \tag{11}$$

where, A is the region of the object in the 3D space, and a is the region that the object projects on the 2D image plane.

From (9) and (10) , we get: $\left|\partial(x, y)/\partial(X, Y)\right| = \begin{vmatrix} \partial x/\partial X & \partial x/\partial Y \\ \partial y/\partial Y & \partial y/\partial Y \end{vmatrix} = \frac{\lambda^2 f_x f_y}{Z^2} = \left|\Lambda\right|$. So:

$$M_{00} = \left|\Lambda\right| \iint_{(X,Y)\in A} dXdY \tag{12}$$

From (9) we get the equation of (13):

$$\left[X, Y\right]^T = (1/s)\Theta^{-1}\Lambda^{-1}\left[x', y'\right]^T \tag{13}$$

As

$$\iint_{(X,Y)\in A} dXdY = \iint_{(x',y')\in a'} \left|\partial(X, Y)/\partial(x', y')\right| dx'\, dy' \tag{14}$$

and the following equation obtained from (13) and (10):

$$\left|\partial(X, Y)/\partial(x', y')\right| = \left(1/s\right)^2 \left|\Theta^{-1}\right| \left|\Lambda^{-1}\right| \tag{15}$$

In (14), a' is the projection region of the object after motion. Thus, (16) can be deduced from (14) and (15):

$$\iint_{(X,Y)\in A} dXdY = \left(1/s\right)^2 \left|\Theta^{-1}\right| \left|\Lambda^{-1}\right| \iint_{(x',y')\in a'} dx'\, dy' \tag{16}$$

From (10), we know: $\left|\Theta^{-1}\right| = \left|\Theta\right| = 1$.

So, (17) is established from (12) ,(16), and (2):

$$M_{00} = (1/s^2)\left|\Lambda\right| \left|\Lambda^{-1}\right| M_{00}' = \left(1/s^2\right) M_{00}' \tag{17}$$

Then, from (10) and (17), the equation of (8) is gotten and the lemma is proved.

2.3 The Relationship between 2nd-Order Image Central Moments and Object Orientation

As described in Fig.2, in the frame {C} M is let to be the gravity center of the object, P' and P be the position of an arbitrary point of the object after and before motion. In the frame {I}, the points of m, p, p' are respectively let to be the corresponding point at which the point of M, P, and P' projects in the image.

If an object is plan and always perpendicular to the camera optic axis (Z-axis), then the object is always parallel to the image plane. Then, it is easy to judge that the two triangle of $\Delta MPP'$ and $\Delta mpp'$ are similar. Therefore, the angle of $\angle MPP'$ is equal to the angle of $\angle mpp'$. Thus, we can draw a conclusion of the following:

If an object is always perpendicular to the camera optic axis (Z-axis), the object orientation θ is equal to the rotation angle of the region where the object projects on the image.

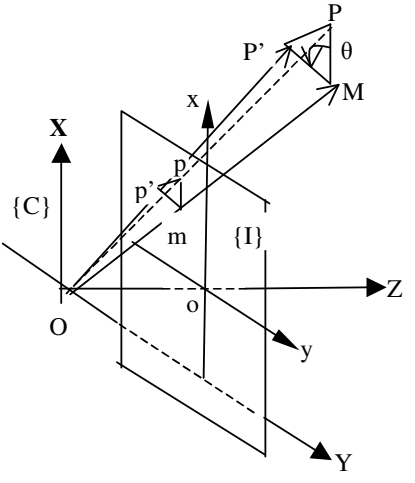


Fig. 2. The object orientation of θ in the frame {C} and the region rotation angle in the frame {I}

In the references (6), we found that for any region on the image, its rotation angle of σ can be estimated by the image central moments of order 2:

$$\sigma = \frac{1}{2}\arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right)$$

So, if the object is always perpendicular to the Z-axis, the object orientation θ can be estimated from σ , that is:

$$\theta = \frac{1}{2}\arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \tag{18}$$

3 Image Feature Selection

If Z' is equal to the given expectation Z_d and the zero-order geometric moment of the corresponding image is denoted by $(M_{00})_d$, then (8) can be written as:

$$(Z_d / Z)^2 = M_{00} / (M_{00})_d \tag{19}$$

When the camera moves to the position where the object depth Z is equal to Z_d, from (19) it is clearly to see that M_{00} is equal to $(M_{00})_d$. At this time, the task of object depth control can be finished. So, we select the zero-order image geometric moment as the image feature reflecting object depth. The error of e_{depth} is:

$$e_{depth}(t) = (M_{00})_d - M_{00}(t) \tag{20}$$

Let $\{(\mu_{11})_d, (\mu_{20})_d, (\mu_{02})_d\}$ denote the set of the 2nd-order central moments of the ideal image which is taken at the time when the object orientation takes its ideal value θ_d. If $\{\mu_{11}, \mu_{20}, \mu_{02}\}$ is equal to $\{(\mu_{11})_d, (\mu_{20})_d, (\mu_{02})_d\}$, then from (18), we know that the object orientation θ is equal to θ_d, the task of object orientation control is finished. So, we select the set of 2nd-order image central moments of $\{\mu_{11}, \mu_{20}, \mu_{02}\}$ as the image features reflecting the object orientation.

In summary, the set of $\{M_{00}, \mu_{11}, \mu_{20}, \mu_{02}\}$ is selected as the image feature set in our visual servoing system, in which M_{00} is used to control object depth, and the set of $\{\mu_{11}, \mu_{20}, \mu_{02}\}$ is used to control object orientation.

4 Control Structure of Visual Servoing System

A hybrid visual servoing system is designed based on the selected image moments. The control structure of the system is described in Fig.3.

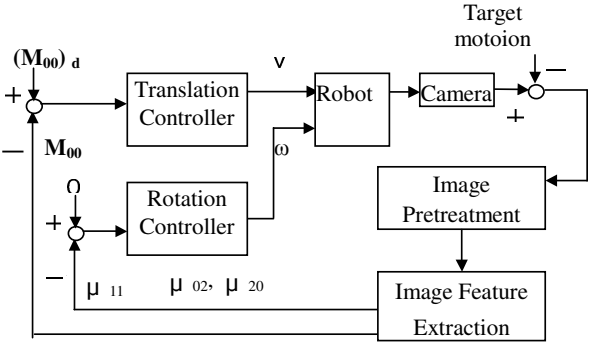


Fig. 3. The control structure of our hybrid servoing system

In the proceeding of control, the object is static and the camera is driven by the controllers to move to its ideal position where the object depth is equal to the given expectation and the object orientation is equal to zero (i.e., θ_d=0).

From (18), we know if μ_{11} is equal to zero, then θ must be zero. So we set the error of e_θ to reflect the error of the object orientation:

$$e_\theta(t) = (\mu_{11})_d - \mu_{11}(t), \quad (\mu_{11})_d = 0 \tag{21}$$

According to (20) and (21), the state vector of our system can be selected as:

$$X=[\,M_{00}, \mu_{11}]^T$$

Then: the desired state vector $X^*=[(M_{00})_d, 0]$, the system error $E=X^*-X$.

When $|E|<\varepsilon$, Camera stops.

As description in the Fig3, our visual servoing controller consists of two parts: one is a translation controller which treats the camera moving translationally along

the Z-axis, the other is a rotation controller which treats the camera rotating around the Z-axis. The control signal of U consists of two components: the linear velocity and the rotation speed of the camera in the Z-axis direction:

$$U = [u_1, u_2]^T = [v_{cZ}, \omega_{cZ}]^T \tag{22}$$

The Translation controller in our visual servoing system is designed as a fuzzy logical controller, which has been introduced in the reference (7) in detail. The rotation controller will be designed in the next section.

5 Rotation Controller Design

Let e_θ and ω_{cz} be the input and output of the rotation controller. An adaptive visual controller is then designed by a Lyapunov function to make the system asymptotically stable. The Lyapunov function L is selected as: $L = pe_\theta^{\ 2}/2$.

Let p>0, then the derivative of L is:

$$\dot{L} = pe_\theta \dot{e}_\theta, \ \text{p>0} \tag{23}$$

From (21), the derivative of e_θ can be written as: $\dot{e}_\theta = -\dot{\mu}_{11}$.

When the above expression is substituted in (23), (23) is changed to (24):

$$\dot{L} = -pe_\theta \dot{\mu}_{11}, \ \text{p>0} \tag{24}$$

It is well known that if $\dot{L} < 0$, the system can be asymptotically stable. Because p> 0, to ensure $\dot{L} < 0$ we have:

$$\dot{\mu}_{11} = k_u e_\theta, \quad \text{k}_\text{u} > 0 \tag{25}$$

In the reference (11), we found that :

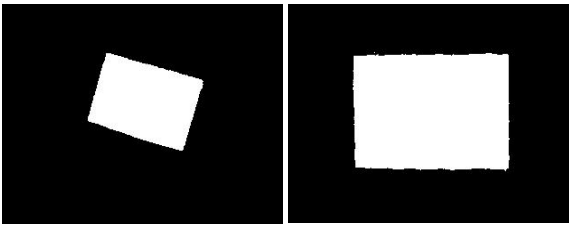
$$\dot{\mu}_{11} = (4\mu_{11} v_{cZ}/Z) + (\mu_{02} - \mu_{20})\omega_{cZ} \tag{26}$$

Considering (19), (26) and (27), we design our rotation controller as:

$$\omega_{cz} = \frac{k_u e_u - (C_d \mu_{11}\sqrt{M_{00}})v_{cz}}{\mu_{02} - \mu_{20}}, \ C_d = (Z_d \sqrt{(M_{00})_d})^{-1}, \ k_u > 0 \tag{27}$$

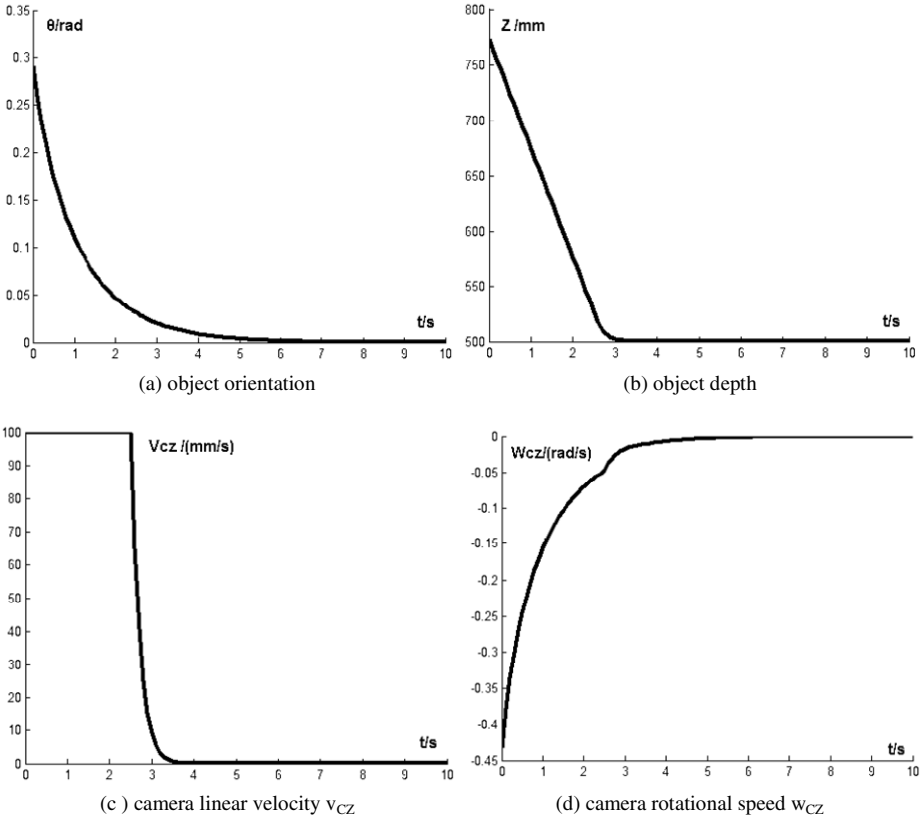
6 Simulation Results

The initial object image is displayed in Fig.4(a), which is taken at the initial position where object depth Z is 760mm and object orientation θ is 0.3radian. The ideal object image is displayed in Fig.4(b), which is taken at the ideal position where the ideal object depth Z_d is 500mm and the ideal object orientation θ_d is 0radian.



(a) initial object image (b) ideal object image

Fig. 4. Object images



(a) object orientation (b) object depth

(c) camera linear velocity v_{CZ} (d) camera rotational speed w_{CZ}

Fig. 5. Simulation results of translation controller and rotational controller

The image moments of the above two images can be extracted by the module of image feature extraction described in Fig.3. Their values are:

(1) the ideal image moment $(M_{00})_d$=21717; (2) the initial image moments:

$$M_{00}(0) = 9090,\ \mu_{11}(0) = -1595000,\ \mu_{20}(0) = 57851000,\ \mu_{02}(0) = 28252000.$$

Supposed that:

(a) camera parameter: λ =16mm, k_x=40.25pixels/mm, k_y=37.21pixels/mm.
(b) sampling time: T_s=100ms.
(c) for Rotation controller: k_u=0.8

The simulation results of our hybrid servoing system in control of the rotation controller designed by (27) and the translation controller introduced in the reference (7) can be seen in Fig.5.

References

1. Hutchinson, S., Hager, G.D., Corke, P.I.: A tutorial on visual servo control. IEEE Trans. Robot. Automation 12, 651–670 (1996)
2. Espiau, B., Chaumette, F., Rives, P.: A new approach to visual servoing in robotics. IEEE Trans. Robot. Automation 8, 313–326 (1992)
3. Cork, P., Hutchinson, S.: A new partitioned approach to image-based visual servo control. IEEE Trans. Robot. Automation 17, 507–515 (2001)
4. Malis, E.: 2-1/2-D visual servoing. IEEE Trans. Robot. Automation 15, 238–250 (1999)
5. Deguchi, K.: Optimal motion control for image-based visual servoing by decoupling translation and rotation. In: Int. Conf. Intelligent Robots and Systems, pp. 705–711. IEEE Press, Victoria (1998)
6. Wohn, K., Wu, J.E.: Estimation the finite displacement using moments. Pattern Recognition Letters 11, 371–378 (1990)
7. Shen, X.J.: Fuzzy control for a visual servoing system. In: The 2009 WRI Global Congress on Intelligent System, vol. 3, pp. 93–97. IEEE Press, XiaMen (2009)
8. Lo, C., Don, H.: 3D moment forms: their construction and application to object identification and position. IEEE Transaction on Pattern Analysis and Machine Intelligence 11, 1053–1064 (1989)

A Novel Motion Detection Approach for Large FOV Cameras*

Hongfei Yu, Wei Liu, Bobo Duan, Huai Yuan, and Hong Zhao

Software Center of Northeastern University,
110004 Shenyang, China
{yuhf,lwei,duanbb,yuanh,zhaoh}@neusoft.com

Abstract. Moving objects detection by an also moving camera plays an important role in driver assistance systems and robot navigations. Many motion detection methods have been proposed until now. But most of them are based on normal cameras with a limited view and not suitable for large FOV (field of view) cameras. For motion detection using large FOV cameras, there are two main challenges. One comes from difficulties to tell moving objects from moving background due to camera motion. The other comes from the image distortion brought by large FOV cameras. These two problems are solved in our approach by a novel motion detector which can be considered as a special motion constraint based on virtual planes. The experimental results under various scenes illustrate the effectiveness of this work.

Keywords: moving detection; large FOV; motion constraint; virtual planes.

1 Introduction

Moving objects detection has long been being an important theme in computer vision. It can be used in many applications such as driver assistance, road safety, surveillance and robot navigation. In the past, normal cameras with a limited field of view were generally used for motion detection and numerous algorithms were proposed to tackle this issue. In recent years, large FOV cameras, such as fish-eye camera and omnidirectional camera, are more and more popular as their much wider imaging angle [1]. Many researchers have focused on the research of driver assistance systems based on large FOV cameras in order to avoid the blind area [2]. So motion detection algorithm based on a large FOV camera has important significance.

There are two main challenges in motion detection using moving large FOV cameras. Firstly, compared with using camera which stays at a fixed location, it is more difficult, as the moving camera also induces apparent motion of the scene. Secondly, the camera's large angle of view brings image distortion of object's real shape, which makes many feature based algorithms inapplicable.

* This work was supported by national high technology research and development program 863 under grant No.2009AA011908.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 311–320, 2010.

In this paper, we solve the above two questions by setting six special virtual planes and a novel motion detector is designed based on these virtual planes. It is suitable for road scene and can be used in driving assistance systems.

This paper is organized as follows. Section 2 gives a briefly review of existing approaches related to our work. Section 3 introduces the large FOV camera model which can be considered as an extension to conventional pine-hole model. Section 4 gives a detailed description to our approach. Section 5 shows the experimental results under various scenes captured by a fish-eye camera. Finally, the conclusion is given in section 6.

2 Related Works

Numerous methods have been proposed to tackle the issue of motion detection. When the camera stays at a fixed location, typically change detection methods are used such as inter-frame difference and background modeling. After a little modification these methods are also suitable for large FOV cameras [3]. When the camera is moving, the task becomes complex due to the moving background. The commonly used motion detection methods in this case can be classified into methods based on motion compensation [4] [5] [6], optical flow [7] [8] [9] and geometric constraint [10] [11] [12].

Motion compensation method adopts the idea of firstly compensating the global motion and then finding the motion regions by frame difference between current frame and the compensated frame. This method is reliable and robust when applied to distant scenes, planar scenes, or the camera is undergoing only rotations and zooms. In [6], a moving human detection approach based on an omni-directional camera is proposed. The author uses motion compensation method to get the potential human regions on the assumption that the background is relatively far away from the camera and roughly in the same plane. However, this assumption is difficult to meet in road scenes when utilizing large FOV cameras, as they broaden the field of view and many nearby objects are mapped into image which brings strong parallax.

Optical flow method is a classical approach to analyze image apparent motion. The continuity of optical flow fields is usually used in moving object detection and segmentation. But determining accurate optical flow is a difficult and computationally expensive operation [7]. In the image that shot by large FOV camera, since the imaging distortions complicate the optical flow field, it is more difficult to detect moving objects from such optical flow field. In [9], an approaching vehicle detection algorithm using in-vehicle fish-eye cameras is proposed. In this paper, the fish-eye images are firstly rectified into plane images and then motion region is detected by analyzing the optical flow field in each rectified images. However the rectification process will cause the imaging angle loss. Although rectifying the image to more planes can avoid this, the processing time will prolong and the image pixel accuracy will decrease.

In geometric constraint method, there are two-view methods and n-view methods. The basic idea of this kind of methods is to utilize various geometric constraints to detect moving objects. The epipolar constraint [10] is a usually used geometric constraint for motion detection in two views. However, there exists degenerated case [11]. In order to solve this problem, n-view methods are developed such as trilinear constraint method [12]. But computing the trifocal tensor is a difficult work.

Besides the above methods some special motion analysis methods have been developed for large FOV cameras [13] [14]. In [13], a framework for performing stereo analysis using a pair of vehicle mounted omnidirectional cameras is described. Although using more than one large FOV cameras can get the depth information of the scene and realize motion detection indirectly, the stereo matching is somewhat time-consuming and sensitive to the accuracy and stability of the calibration.

The algorithm proposed in this paper can detect moving objects using only one large FOV camera. It detects moving objects in a direct way and doesn't need to get the depth information. It can be seen as a special kind of geometric constraint method.

3 Large FOV Camera Model

The imaging model of normal camera is based on pinhole camera model, while the imaging model of large FOV camera can be considered as spherical projection model. For comparison the two kinds of camera models are shown in Fig. 1, where $O-XYZ$ is the camera coordinate system, P is a space point, and its projection point in the image is p. Suppose the coordinate of P in $O-XYZ$ is (X,Y,Z). For normal cameras, the normalization vector p_s of P can be computed by the equation $p_s=(\frac{X}{Z},\frac{Y}{Z},1)$, but it's not suitable for large FOV cameras. Thus, we adopt the spherical projection method in [14] to define the normalization vector p_s of P in large FOV cameras.

$$p_s=\frac{1}{\sqrt{X^2+Y^2+Z^2}}(X,Y,Z)\cdot \tag{1}$$

Suppose the image coordinate of p is (u,v). A one-to-one projection from p to P can be given as

$$p(u,v)=M(p_s(x_s,y_s,z_s))\cdot$$

Where M is a mapping rule which can be deduced by the image calibration algorithm in [15], so

$$p_s(x_s,y_s,z_s)=M^{-1}(p(u,v))\cdot \tag{2}$$

Here, M^{-1} is the reverse mapping of M and $p_s=p_s(x_s,y_s,z_s)$. Equation (1) and (2) give a relationship between p and P by p_s. This relationship will be used frequently in the following discussion.

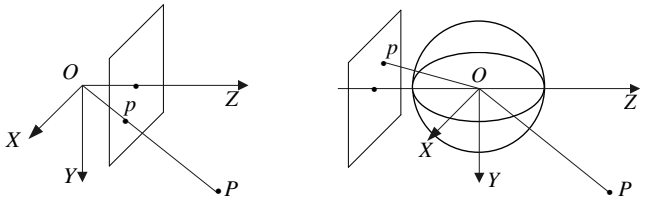


Fig. 1. Imaging models: (a) pin-hole model which is suitable for normal cameras; (b) large FOV camera model

4 Proposed Algorithm

The algorithm is comprised of three main steps: (1) setting virtual planes, in which virtual camera coordinate systems between two consecutive frames are built and six special virtual planes are set around them; (2) computing projection points, in which the correlation between the point in large FOV image and virtual plane is deduced and the virtual plane projection points of image feature points are computed by this correlation; (3) detecting moving points, in which moving points in image are detected according to their displacement vectors in virtual planes.

4.1 Setting Virtual Planes

The motion of camera coordinate system in two consecutive frames is shown in Fig. 2 (a), where $O_{t-1}-X_{t-1}Y_{t-1}Z_{t-1}$ and $O_t-X_tY_tZ_t$ denote the camera coordinate system at time t-1 and time t respectively and α,β,γ denote the rotation angles around X,Y,Z axes respectively. The coordinate of origin O_t in $O_{t-1}-X_{t-1}Y_{t-1}Z_{t-1}$ is (T_x,T_y,T_z). In order to set virtual planes, two virtual camera coordinate systems $O'_{t-1}-X'_{t-1}Y'_{t-1}Z'_{t-1}$ and $O'_t-X'_tY'_tZ'_t$ are set, as is shown in Fig. 2 (b). $O'_{t-1}-X'_{t-1}Y'_{t-1}Z'_{t-1}$ is gotten by rotating $O_{t-1}-X_{t-1}Y_{t-1}Z_{t-1}$ for θ degrees around the vector N. $O'_t-X'_tY'_tZ'_t$ is gotten by firstly rotating $O_t-X_tY_tZ_t$ for $-\gamma,-\beta,-\alpha$ degrees around the Z,Y,X axes in turn, and then rotating for θ degrees around the vector N. Here,

$$\theta = arctg(\frac{\sqrt{T_x^2+T_y^2}}{T_z}),\ N=(T_x,T_y,T_z)\times(0,0,1)\ .$$

After setting virtual coordinate systems in the two consecutive frames, six virtual planes are set in either of them. Let $plane(i), i=1,2,\cdots,6$ denote the virtual planes, where i denotes the plane index. The expressions of these virtual planes are given as $X=1$, $X=-1$, $Y=1$, $Y=-1$, $Z=1$ and $Z=-1$ in turn. Fig. 2 (c) illustrates the appearance of these virtual planes.

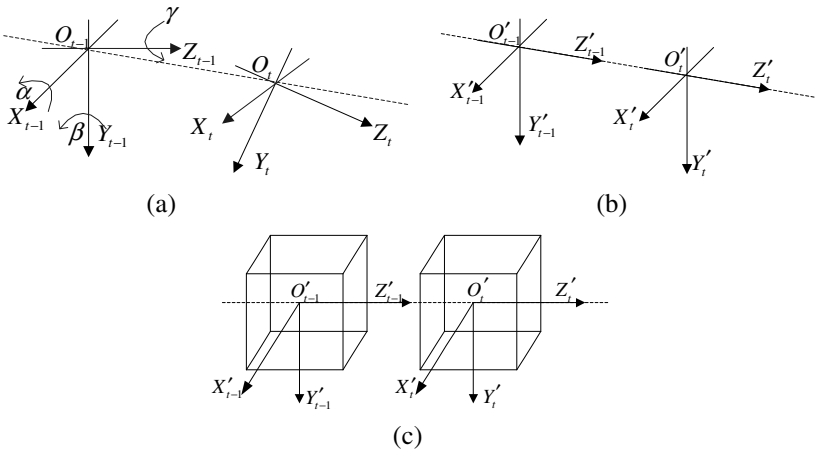


Fig. 2. (a) The camera coordinate systems between consecutive frames. (b) The new coordinate systems which are gotten by rotating the coordinate systems in (a). (c) Foundation of virtual planes in consecutive frames.

4.2 Computing Projection Points

For simplification, only the image feature points are analyzed. To select and match feature points between two frames KLT method [16] is employed because of its accuracy and efficiency.

Given a pair of feature points (p^{t-1}, p^t), where p^{t-1} denotes the feature point in frame t-1 and p^t denotes the corresponding point in frame t, its projection point pair (p_{t-1}, p_t) in virtual planes can be computed by the following steps.

Step1. Compute the normalization points $p_s^{\,t}(x_s^{\,t}, y_s^{\,t}, z_s^{\,t})$ of p^t in $O_t - X_t Y_t Z_t$ and $p_s^{\,t-1}(x_s^{\,t-1}, y_s^{\,t-1}, z_s^{\,t-1})$ of p^{t-1} in $O_{t-1} - X_{t-1} Y_{t-1} Z_{t-1}$ respectively by equation (2).

Step2. Compute the normalization points $p_s^{\,t'}(x_s^{\,t'}, y_s^{\,t'}, z_s^{\,t'})$ of $p_s^{\,t}$ in $O'_t - X'_t Y'_t Z'_t$ and $p_s^{\,t-1'}(x_s^{\,t-1'}, y_s^{\,t-1'}, z_s^{\,t-1'})$ of $p_s^{\,t-1}$ in $O'_{t-1} - X'_{t-1} Y'_{t-1} Z'_{t-1}$ by equation (3).

$$\begin{pmatrix} x_s^{t'} \\ y_s^{t'} \\ z_s^{t'} \end{pmatrix} = R^{-1} r^{-1} \begin{pmatrix} x_s^{t} \\ y_s^{t} \\ z_s^{t} \end{pmatrix}, \begin{pmatrix} x_s^{t-1'} \\ y_s^{t-1'} \\ z_s^{t-1'} \end{pmatrix} = R^{-1} \begin{pmatrix} x_s^{t-1} \\ y_s^{t-1} \\ z_s^{t-1} \end{pmatrix}. \tag{3}$$

Here,

$$r = \begin{pmatrix} c_\gamma c_\beta & c_\gamma s_\beta s_\alpha - s_\gamma c_\alpha & c_\gamma s_\beta c_\alpha + s_\gamma s_\alpha \\ s_\gamma c_\beta & s_\gamma s_\beta s_\alpha + c_\gamma c_\alpha & s_\gamma s_\beta c_\alpha - c_\gamma s_\alpha \\ -s_\beta & c_\beta s_\alpha & c_\beta c_\alpha \end{pmatrix},$$

$$R=\begin{pmatrix} n_1^2+(1-n_1^2)c_\theta & n_1n_2(1-c_\theta)-n_3s_\theta & n_1n_3(1-c_\theta)+n_2s_\theta \\ n_1n_2(1-c_\theta)+n_3s_\theta & n_2^2+(1-n_2^2)c_\theta & n_2n_3(1-c_\theta)+n_1s_\theta \\ n_1n_3(1-c_\theta)-n_2s_\theta & n_2n_3(1-c_\theta)+n_1s_\theta & n_3^2+(1-n_3^2)c_\theta \end{pmatrix},$$

$$(n_1,n_2,n_3)=\frac{(T_x,T_y,T_z)\times(0,0,1)}{\left|(T_x,T_y,T_z)\times(0,0,1)\right|}.$$

Where $s_x \equiv \sin x$, $c_x \equiv \cos x$.

Step3. Compute the virtual plane projection points $p_t(x_t', y_t', z_t')$ of ${p_s}^{t'}$ in $O_t'-X_t'Y_t'Z_t'$ and $p_{t-1}(x_{t-1}', y_{t-1}', z_{t-1}')$ of ${p_s}^{t-1'}$ in $O_{t-1}'-X_{t-1}'Y_{t-1}'Z_{t-1}'$ respectively. The virtual plane which p^t is projected into is determined by the direction of vector $O_t'{p_s}^{t'}$. Here, we choose the one whose projection point has the smallest Euclidean distance from origin O_t'. As for ${p_s}^{t-1'}$, it will be projected into the virtual plane which has the same plane index as that of ${p_s}^{t'}$.

4.3 Detecting Moving Points

Feature points p^t and p^{t-1} can be seen as images of a space point P in different time. If P is still, the normalization point $p_s{}^t(x_s{}^t, y_s{}^t, z_s{}^t)$ and $p_s{}^{t-1}(x_s{}^{t-1}, y_s{}^{t-1}, z_s{}^{t-1})$ will meet the following equation:

$$\begin{pmatrix} x_s{}^t \\ y_s{}^t \\ z_s{}^t \end{pmatrix} = r\begin{pmatrix} x_s{}^{t-1} \\ y_s{}^{t-1} \\ z_s{}^{t-1} \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix}. \tag{4}$$

From equation (3) and (4), we have:

$$\begin{pmatrix} x_s^{t'} \\ y_s^{t'} \\ z_s^{t'} \end{pmatrix} = \begin{pmatrix} x_s^{t-1'} \\ y_s^{t-1'} \\ z_s^{t-1'} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ T_z' \end{pmatrix}. \tag{5}$$

Where,

$$T_z' = \sqrt{T_x^2+T_y^2+T_z^2}.$$

We can know from equation (5) that the relation between p^t and p^{t-1} is simplified by using their virtual plane projection coordinates. When $i=1,\cdots,4$, $plane(i)$ is parallel to line $O_{t-1}'O_t'$, so the displacement vector $d(x_t'-x_{t-1}', y_t'-y_{t-1}', z_t'-z_{t-1}')$ of p^t in $plane(i)$ is also parallel to line $O_{t-1}'O_t'$. When $i=5,6$, $plane(i)$ is perpendicular to line $O_{t-1}'O_t'$, so the displacement vector $d(x_t'-x_{t-1}', y_t'-y_{t-1}', z_t'-z_{t-1}')$ of p^t in $plane(i)$ is

perpendicular to $O'_{t-1}O'_t$ and intersects with $O'_{t-1}O'_t$. Thus, a motion constraint of a still space point P can be given by the direction of its displacement vector d. This constraint can be written as:

$$f_{angle}(d, d_s \cdot sign(T_z)) = 0 \quad . \tag{6}$$

Here,

$$d_s = \begin{cases} (0,0,-\sqrt{T_x^2+T_y^2+T_z^2}), & i=1,\cdots,4 \\ (x'_t, y'_t, 0), & i=5 \\ (-x'_t, -y'_t, 0), & i=6 \end{cases}, \quad sign(T_z) = \begin{cases} 1, & T_z > 0 \\ 0, & T_z = 0 \\ -1, & T_z < 0 \end{cases}.$$

$f_{angle}(d, d_s)$ denotes the angle between vectors d and d_s, and

$$f_{angle}(d, d_s) = \arccos \frac{d \cdot d_s}{|d| \cdot |d_s|} \quad .$$

When P is moving, it will not meet this constraint. So moving points can be detected by equation (6).

5 Experimental Results

Our algorithm is evaluated on real sequences which captured by a fish-eye camera whose field of view is 185 degrees. The size of captured images is 720×480 pixels and the frame frequency is 30 fps. When shooting, the camera is fixed to a car and moves with it and there is no much illumination change. The camera's height is 0.8 meters. The scenes in the sequences can be grouped into planar scenes and strong parallax scenes. In the first group, the background is far from the camera, or the flat road occupies dominant part. In the second group, the background contains large depth variations and this kind of scene generally exists in our daily life.

In the application of our motion detection algorithm, the required ego-motion parameters are retrieved by a vision-based estimation scheme. When judging whether a feature point in current frame is moving or not, if the length of displacement vector of this feature point is less than 0.5 pixels it will be omitted, for the direction of its displacement vector will be affected by the precision of ego-motion estimation greatly.

5.1 Planar Scenes

Fig. 3 (a) shows crossing pedestrian. The background is relatively far from ego vehicle and the flat road occupies dominant part. Ego vehicle is moving in a straight manner and moving objects which need to be detected are marked with red rectangles. The red points in Fig. 3 (b) denote the final detected moving points. We can see that the moving pedestrians are detected successfully.

We compare our method with motion compensation method used in [6]. In that method, the original images are firstly transformed into cylindrical images as shown in Fig. 3 (c) and then the background offset introduced by ego vehicle is compensated

by using an affine transformation model. Finally, the motion objects are detected by temporal differencing. Fig. 3 (d) shows the detection result of motion compensation. We can see that the moving target is highlighted successfully. Both our method and motion compensation method show promising results in planner scenes.

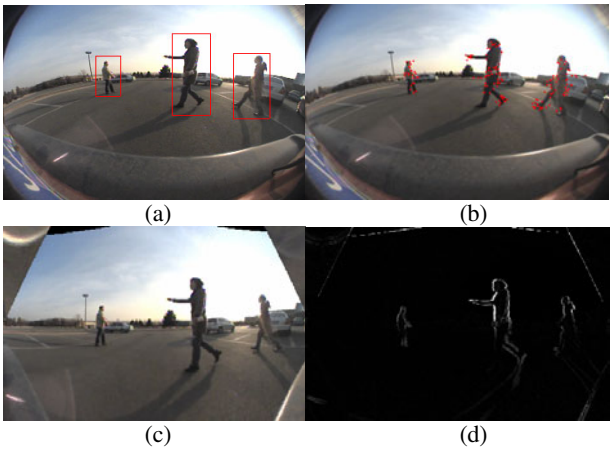


Fig. 3. Motion detection in planar scene: (a) original fish-eye image where the red rectangles mark moving objects need to be detected; (b) motion detection result of our method where the red points denote the detected moving points; (c) cylindrical transformation image of (a); (d) motion detection results of method in [6], where the highlighted regions show the detected moving objects.

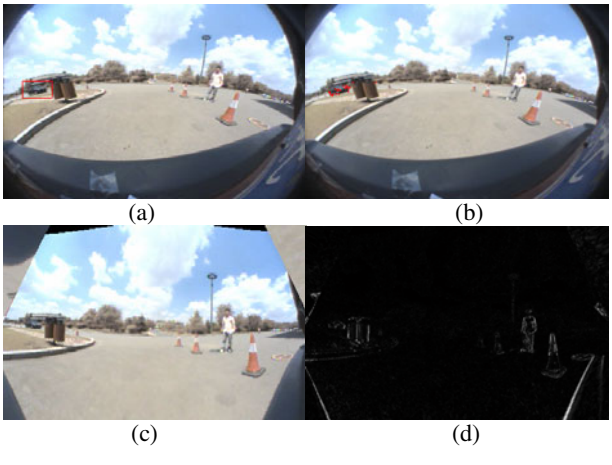


Fig. 4. Motion detection in strong parallax scene: (a) original fish-eye image where the red rectangle marks moving object need to be detected; (b) motion detection result of our method where the red points denote the detected moving points; (c) cylindrical transformation image of (a); (d) motion detection results of method in [6], where the highlighted regions show the detected moving objects.

5.2 Strong Parallax Scenes

A parking lot scene is shown in Fig. 4 (a). The target vehicle is steering near. The static pedestrian, dustbin and road-blocks near ego vehicle induce the strong parallax of the scene. Ego vehicle is moving in a curve manner. Our method can still detect the moving object. But the motion compensation algorithm in [6] doesn't work well as is shown in Fig. 4 (d). Both the moving vehicle and the still obstacles nearby are highlighted. Our method gives a better detection result than motion compensation method when there are strong parallaxes in the scene. More detection results of our algorithm are shown in Fig. 5.

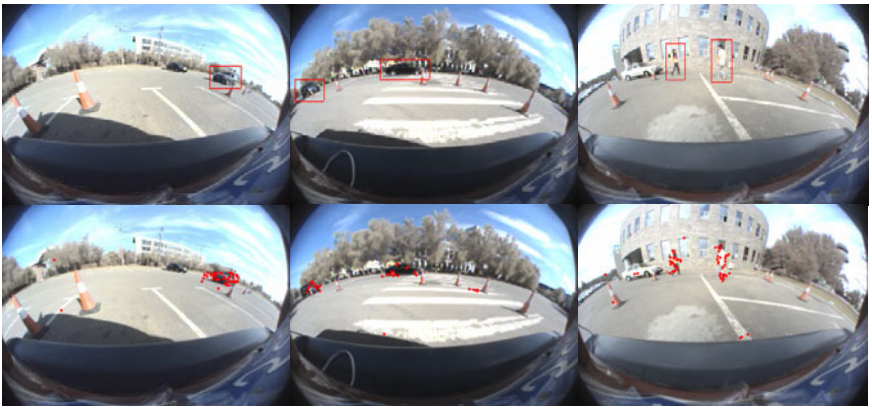


Fig. 5. Experimental results achieved with our approach. The red rectangles in the first row of images mark the moving objects need to be detected. The red points in the second row of images show the moving point detection results.

6 Conclusion

This paper has presented a novel method of moving objects detection which is suitable for large FOV cameras. The proposed method can detect objects both in planar scenes and strong parallax scenes. Up to now, only two consecutive frames are used in order to judge whether a point is moving or not. Since the error matching feature point pairs can cause false detection, a further improvement can be reached by utilizing a more sophisticated tracking method within multi-frames. How to overcome the degradation case of our algorithm is another further work.

References

1. Benosman, R., Kang, S.B. (eds.): Panoramic Vision: Sensors, Theory, and Applications. Springer, Heidelberg (2001)
2. Gandhi, T., Trivedi, M.: Motion based vehicle surround analysis using an omni-directional camera. In: 2004 IEEE Intelligent Vehicles Symposium, pp. 560–565 (2004)

3. So, I.M., Han, D.K., Kang, S.K., Kim, Y.U., Jung, S.T.: Recognition of fainting motion from fish-eye lens camera images. In: The 23rd International Conference on Circuits/ Systems, Computers and Communications (ITC-CSCC 2008), pp. 1205–1208 (2008)
4. Jung, B., Sukhatme, G.S.: Real-time motion tracking from a mobile robot. Technical report CRES-05-008, University of Southern California (2005)
5. Fardi, B., John, T., Wanielik, G.: Non-rigid-motion recognition using a moving mono camera. In: 2009 IEEE Intelligent Vehicles Symposium, pp. 221–226 (2009)
6. Liu, H., Dong, N., Zha, H.: Omni-directional vision based human motion detection for autonomous mobile robots. In: 2005 IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2236–2241 (2005)
7. Krüger, W., Enkelmann, W., Rossle, S.: Real-time estimation and tracking of optical flow vectors for obstacle detection. In: Proceedings of the Intelligent Vehicles 1995 Symposium, pp. 304–309 (1995)
8. Shafie, A.A., Hafiz, F., Ali, M.H.: Motion detection techniques using optical flow. World Academy of Science, Engineering and Technology, 559–561 (2009)
9. Nagata, A., Ozawa, S., Yanagawa, H.: Detecting approaching vehicles from the rear side using in-vehicle fish-eye cameras. In: Papers of technical Meeting on Information Processing, IEE Japan, pp. 65–70 (2007)
10. Yamaguchi, K., Kato, T., Ninomiya, Y.: Vehicle ego-motion estimation and moving object detection using a monocular camera. In: The 18th International Conference on Pattern Recognition (ICPR 2006), vol. (4), pp. 610–613 (2006)
11. Yuan, C., Medioni, G., Kang, J., Cohen, I.: Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(9) (2007)
12. Hartley, R., Vidal, R.: The multibody trifocal tensor: motion segmentation from 3 perspective views. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 769–775 (2004)
13. Gandhi, T., Trivedi, M.M.: Vehicle mounted wide FOV stereo for traffic and pedestrian detection. In: IEEE International Conference on Image Processing, vol. 2, pp. 121–124 (2005)
14. Li, S., Fukumori, K.: Spherical stereo for the construction of immersive VR environment. In: Proceedings of the IEEE virtual reality 2005 (VR 2005), pp. 217–222 (2005)
15. Kannala, J., Brandt, S.S.: A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8) (2006)
16. Tomasi, C., Kanade, T.: Detection and tracking of point features. Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-91-132 (1991)

Large Scale Visual Classification via Learned Dictionaries and Sparse Representation

Zhenyong Fu[1], Hongtao Lu[1], Nan Deng[2], and Nengbin Cai[2]

[1] Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
[2] Forensic center of Shanghai Police
Shanghai, China

Abstract. We address the large scale visual classification problem. The approach is based on sparse and redundant representations over trained dictionaries. The proposed algorithm firstly trains dictionaries using the images of every visual category, one category has one dictionary. In this paper, we choose the K-SVD algorithm to train the visual category dictionary. Given a set of training images from a category, the K-SVD algorithm seeks the dictionary that leads to the best representation for each image in this set, under strict sparsity constraints. For testing images, the traditional classification method under the large scale condition is the k-nearest-neighbor method. And in our method, the category result is through the reconstruction residual using different dictionaries. To get the most effective dictionaries, we explore the large scale image database from the Internet [2] and design experiments on a nearly 1.6 million tiny images on the middle semantic level defined based on WordNet. We compare the image classification performance under different image resolutions and k-nearest-neighbor parameters. The experimental results demonstrate that the proposed algorithm outperforms k-nearest-neighbor in two aspects: 1) the discriminative capability for large scale visual classification task, and 2) the average running time of classifying one image.

Keywords: visual classification, dictionary learning, sparse representation.

1 Introduction

Classifying visual objects into different categories is one of the core problems in computer vision, which has received much attention over the past decades. In recent years, there are two main kinds of visual classification approach: generative [7, 10, 16] and discriminative approaches [8, 9]. They are all benefit from the advance of local appearance descriptors [11, 12, 13, 14]. With the local feature description, the images can be represented by the bag of visual words which are built by quantized the local descriptors, such as SIFT [11]. This greatly simplifies the analysis. The generative models represent the image data by a co-occurrence

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 321–330, 2010.

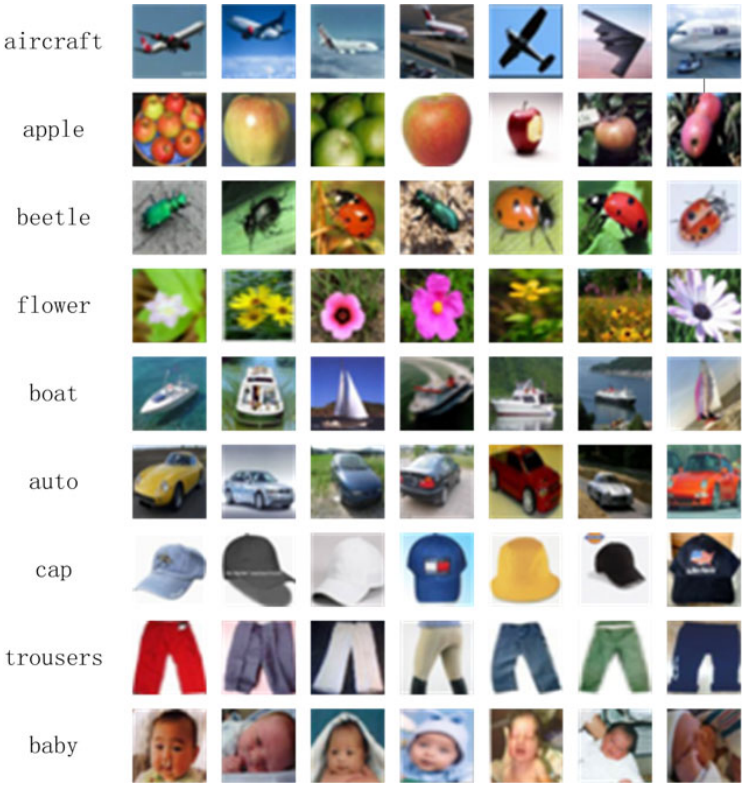


Fig. 1. Image samples from tiny images database. There are nine categories in this figure: aircraft, apple, beetle, flower, boat, auto, cap, trousers and baby. The tiny images database totally contains 1,608,326 color images of size 32×32 pixels, labeled by 53,650 nouns in English, which are a subset of the lexicon of WordNet.

matrix, a table of the counts of each visual word in each document. Then under the bag-of-visual-words representation, the text analysis technique can be adapted to the computer vision area, such as the probabilistic Latent Semantic Analysis (pLSA) [19, 20] and the Latent Dirichlet Allocation (LDA) [18]. The discriminative method use the feature pyramid to construct the effective kernel function, and then train the classifier, such as SVM (support vector machine), to finish the classification task. The above generative and discriminative approaches have some shortcomings: (1) the class counts are very limited in these visual classification algorithms, it usually has less than ten classes in the reported results; (2) the training process of classification model is very slow in the generative and discriminative methods; (3) there usually have many model parameters; (4) the images used to train the classification model are very limited, it is usually several thousands.

With the fast development of the Internet, people can access overwhelming amounts of image data from the Internet. The Internet has become a large

scale distributed database. The tiny image database is one of typical image databases using the large amount of image data in the Internet [2]. Other such large scale image databases include the ImageNet [21] and LabelMe [22]. The construction processes of these large scale image databases are very similar each other. They are all built upon the backbone of the WordNet structure [3], which reflects the semantic relationship of English language. There are total about 80,000 synsets in WordNet. These large-scale image databases use the synsets of WordNet as query words of search engines, such as Google, Bing, and Yahoo, to collect images from searching results. The collected images are organized using the hierarchical structure of WordNet. The data count of these large-scale image databases usually beyond one million. The simplified tiny images database totally contains 1,608,326 color images of size 32×32 pixels, labeled by 53,650 nouns in English, which are a subset of the lexicon of WordNet. Figure 1 shows the image samples of nine categories from tiny images database. They are aircraft, apple, beetle, flower, boat, auto, cap, trousers and baby. In this paper, we use the images from tiny images database as our experimental data. The large scale images database provides the following benefits: (1) the potential visual categories count are vast; (2) the training images is a dense subset of the total natural images; (3) the hierarchical structure of the large scale images database provides the natural different semantic scales. The above three points are very favorable for the large scale visual classification task.

Recent years, the sparse representations of signals have attracted a growing interest in the search area of information theory, signal processing, image processing and computer vision owing to its powerful discriminative and reconstructive ability [1, 4, 5, 6], which can be used to the visual classification task. Using an overcomplete dictionary matrix that contains prototype signal-atoms for columns, a signal can be represented as a sparse linear combination of these atoms. The reconstructive representation of the original signal may either be exact or approximate. Obtaining an overcomplete dictionary from a set of signals is the core problem of sparse representation which allows us to represent signals as a sparse linear combination of dictionary atoms. Pursuit algorithms are then used for signal decomposition to get sparse representation. The K-SVD is a novel method for training overcomplete dictionaries that lead to sparse signal representation. It has been used in many research area of image processing and computer vision, such as image denoising, super-resolution and image and video compression.

In this paper, we use the sparse representation to solve the large scale image classification problem. We firstly choose the middle semantic level in WordNet as the target concept. The images from the hyponyms of each target concept which include the target concept itself are collected as the training images. Then we use the K-SVD algorithm to learn the dictionaries using the training images. For the testing image, we use the dictionaries from different target concepts to compute the sparse representation. The difference between the reconstructive image and the original image is used as the image classification standard, i.e. we classify the testing image as the target concept with the smallest reconstructive

residual. In the experimental part, we compare the classification performance of the proposed method with the k-nearest-neighbor that is the previous method under the large scale visual classification [2].

In section 2, we describe the details of the proposed method. To explain and compare performance, in section 3 we apply the proposed algorithm to the large scale visual classification task. We give experimental results about the classification performance and the average running time of visual classification. We conclude our work in section 4.

2 Visual Classification Based on Sparse Reconstruction

In our work, the natural image $I_{\sqrt{m}\times\sqrt{m}}$ with $\sqrt{m}\times\sqrt{m}$ resolution is represented by a column vector $x \in R^m$, with the m dimension. Assume that there are N natural images in the training set, represented as a data matrix $X = [x_1, x_2, ..., x_N]$. The class label of the image x_i is assumed to be $l_i \in \{1, 2, ..., C\}$, where C is the total number of visual classes. x^t denotes the testing image.

2.1 Image Classification via k-Nearest-Neighbor

The traditional method that performs the large scale visual category task is the k-nearest-neighbor approach [2], which uses the sum of squared differences (SSD) to compare two images. The SSD between two images I_1 and I_2 is:

$$D^2_{ssd} = \sum_{x,y,c} (I_1(x, y, c) - I_2(x, y, c))^2$$

With a fixed k, this method first retrieves k nearest neighbors to the query image from the training set, and then computes the label histogram and sets the label of query image as the most frequent label in these k nearest neighbors.

2.2 Sparse Representation of Signals

We begin the presentation of our proposed large scale classification algorithm by first introducing the sparse representation problem of signals. For the effective sparse representation, we need to construct a dictionary (matrix) $D \in R^{m\times k}$ (where $k > m$ implying that it is redundant). If we have had a dictionary (the above matrix D is known), the sparse representation suggests that every tiny image signal, x, could be represented sparsely over this dictionary. To be more formal, the sparsest representation is the solution of

$$\hat{w} = \underset{w}{\operatorname{argmin}} \| w \|_0 \text{ subject to } \| Dw - x \|_2 \leq \varepsilon \tag{1}$$

where $\| \cdot \|_0$ is the l^0 norm, counting the nonzero entries of a vector. It means that every signal instance can be represented as a linear combination of few atoms from the redundant dictionary D.

For the overcomplete dictionary D that leads to sparse representations, there are two main types of dictionaries. One is the predefined dictionary, such as overcomplete wavelets, curvelets, contourlets, steerable wavelet filters, Fourier transforms etc.; and the other is the learned dictionary which is trained according to the signals need to be represent. The researches in recent years demonstrate that designing dictionaries based on learning outperform the predefined dictionaries [1, 4, 6]. So, in this work, we use the learned dictionary to achieve the visual classification tasks.

2.3 Visual Dictionary Learning via K-SVD

We now describe our scheme for image classification problem. Our approach contains two main processes: an off-line K-SVD dictionaries learning process [1] and an online image classifying process. The K-SVD learning process is an off-line procedure, which produces a set of K-SVD image dictionaries that are then considered fixed for the image classifying step. In our method, a single dictionary is learned by the total images from the same category images of the tiny image dataset.

For images set $X = [x_1, x_2, ..., x_N]$ with an unknown dictionary, the dictionary learning and sparse representation problem should be the solution of

$$\hat{D}, \hat{W} = \underset{D,W}{\operatorname{argmin}} \{ \| DW - X \|_F^2 \} \text{ subject to } \forall i, \| w_i \|_0 \leq T_0 \tag{2}$$

where X is the data matrix, D is the dictionary should be learned and W is the sparse representation matrix of X. The K-SVD algorithm minimizes the objective expression in (2) iteratively, which includes the sparse coding stage and codebook update stage. In the sparse coding stage, the algorithm fixes the dictionary D and update the signal coding W; in the codebook update stage, it fixes the W and solved the new dictionary to minimize the objective function.

In the sparse coding stage, the traditional K-SVD algorithm uses the *orthogonal matching pursuit* (OMP) as the sparse signal decomposition solution [23, 24]. However the *orthogonal matching pursuit* is based the l_0 norm minimization, and so is NP-hard and difficult even to approximate. Recent development of the sparse representation theory [25, 26] reveals that if the signal is sparse enough, the solution of the l_0-minimization problem (1) or (2) is equal to the solution to the l_1-minimization problem:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \| w \|_1 \text{ subject to } \| Dw - x \|_2 \leq \varepsilon \tag{3}$$

This is a convex optimization problem and can be solved in polynomial time by standard linear programming methods [27].

In the codebook update stage, the algorithm update one column of dictionary D each time. For each column $d_i \in D$, the algorithm firstly computes the representation error matrix of d_i, which denotes by E_i and is the reconstruction error when remove the d_i from D:

$$E_k = X - \sum_{j \neq i} d_j w_T^j \tag{4}$$

where w_T^j is the jth row in W, the coefficients that correspond to the jth column d_j in dictionary D. Then according to the partial reconstruction property of SVD, the algorithm applies the SVD decomposition to E_k, $E_k = U\Delta V^T$ and get the best rank-1 approximation of E_k in the sense of the Frobenius norm (i.e. the best least-squares approximation) by only using the first column of U, the first column of V and the item $\Delta(1,1)$. So updated column of d_i is also the first column of U.

The above two stages are repeated until convergence to learn the dictionary. After the dictionary learning process, we get a set of visual dictionaries, $D_1 \sim D_C$, and C represents the total count of visual categories.

2.4 Visual Classifier Based on Reconstruction Residual

For more effective visual classification, we combine the visual categories $D_1 \sim D_C$ to one global visual dictionary D. For a testing image x_t, the l_1-minimization algorithm is firstly used for signal decomposition to get sparse representation of x_t with D, and we can get coefficients of sparse representation, w_t. And the, we decompose w_t into the sub coefficient vector, $w_t^1 \sim w_t^C$, according to different categories. Each sub vector of w_t^i accords to each sub matrix of D_i. We use each part of pairs $\{D_i, w_t^i\}$ to reconstruct the testing image x_t. Finally the difference between the reconstructive image and the original image is used as the standard of the image classification. We classify the testing image as the target concept with the smallest reconstructive residual. The sparse reconstruction residual are computed with Frobenius norm, defined as $\| X \|_F = \sqrt{\sum_{ij} X_{ij}^2}$. The final classifying result of the testing image x_t is computed as:

$$l_t^c = \underset{i}{\operatorname{argmin}} \| D_i w_t^i - x_t \|_F^2 \tag{5}$$

where l_t^c represents the final class label of testing image x_t.

3 Experiments and Results

In our experiment, we learn the visual dictionaries with different image resolution and color depth. We will describe the experiment details and model parameters in this section. Before the learning process for each tiny image, the mean image of the examples in one special image category is calculated and subtracted from all the samples of this image category.

We test our image classification approach via learned dictionaries and sparse representation using the tiny images dataset. There are millions of tiny images in our experiment. The tiny image dataset is hierarchically constructed according to the semantic hierarchy of WordNet and all images are crawled from the Web. We randomly selected 5000 images from the tiny images dataset as testing images.

In our experiment, we use the WordNet as a semantic tool to divide the images into different semantics. If we only consider the hypernym and hyponym relationship between different semantics (synset in WordNet), the relationship

between semantics reduce to the 'IS-A' relationship between visual objects and the semantics form a semantic tree. We choose the 'entity' as the root of the sub hierarchy in WordNet which we choose as the experiment semantic set. To the classification problem, too higher and too lower semantic levels are all nonsense. So we choose the middle semantic level which is the eighth level from 'entity' root in the semantic tree. Each semantic in eighth level represents the root of a new sub semantic tree according to the 'IS-A' relationship between semantics and we collect all images from this sub semantic tree (including the root) as the training samples of this semantic. There are many semantics in this level, but not all of them are proper in the learning process. Because we need to learn the overcomplete dictionaries of each visual category, the training samples should be more than the dimension of the atoms in dictionaries. In our experiment, the training samples in each category must be more than the twice of the atom dimension in dictionaries. In our experiment, we test the three kinds of combination of image resolution and color depth, and they are 16×16 resolution with 1 byte depth gray, 16×16 resolution with 3 bytes depth color and 32×32 resolution with 1 byte depth gray. For example, if the training tiny images are 16×16 resolution and gray, the atom dimension is $16 \times 16 \times 1 = 256$, so the training examples must be more than $256 \times 2 = 512$. It is similar that the training samples for 16×16 resolution and color images must be more than $16 \times 16 \times 3 \times 2 = 1536$ and for 32×32 resolution and gray images must be more than $32 \times 32 \times 1 \times 2 = 2048$. The categories that satisfy the above conditions are respectively 222, 67 and 51 in the above three kinds dimension. The total semantics count, for example in the 222 sub semantic trees, are more than 15000, so the potential visual categories count in this paper is also more than 15000. Figure 2 shows the learned dictionaries of the visual concept 'aircraft' with 16×16 gray images.

We compare the image classification performance of our proposed method and the traditional k-nearest-neighbor method under the large scale visual classification task [2]. We summarize the classification result and time consuming at

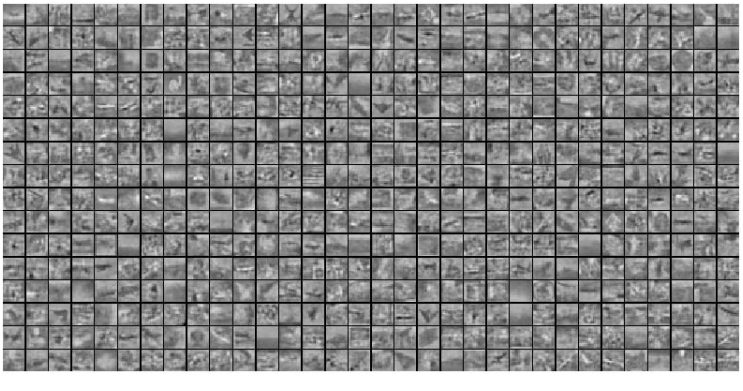


Fig. 2. Learned dictionaries of the visual aircraft category, 16×16 resolution and gray

Table 1. Classification Performance

	Correct Rate	Average Time (second)
SR (16×16 gray)	52.4%	1.58
SR (16×16 color)	53.12%	6.13
SR (32×32 gray)	54.42%	9.49
k-NN ($k = 1$)	13.3%	2413.8
k-NN ($k = 10$)	26.5%	2483.8
k-NN ($k = 50$)	32.2%	2510.6
k-NN ($k = 100$)	31.8%	2403.9
k-NN ($k = 200$)	27.2%	2435.4

table 1. Table 1 shows that SR (sparse representation) methods all outperform k-nearest-neighbor methods in classification correction rate. At the time, the average classifying time of one image in SR method is extremely less than the k-nearest-neighbor time.

4 Conclusion and Future Works

4.1 Conclusions

In this paper, we have investigated ways to classify the large-scale tiny visual images via the learned dictionaries and sparse representation. Our method includes the offline process of learning dictionaries and the classification process using the learned dictionaries. In experiment part, we compare our method based on sparse representation with the k-nearest-neighbor method. The experimental results show that the proposed method outperforms the k-nearest-neighbor method both in the classification performance and in the time consuming.

4.2 Future Works

The proposed method of image classification via the learned dictionaries and sparse representation can be widely used in image understanding related areas, such as image annotation, natural scene recognition, and content-based image retrieval. Our future works will concentrate on these potential application areas.

In the tiny images dataset, almost every image category contains a lot of noise data. The noise in the tiny images dataset will decrease the classification performance very much. So another future research direction will focus on the incremental dictionary learning with which we can use the initial dictionary to reject the noise category images and accept the non-noise images to refine the category visual dictionary. We believe that we can get more robust visual image dictionaries through the dictionary learning process based on incremental manner.

Acknowledgment

This work was supported by the National High Technology Research and Development Program of China (No.2008AA02Z310) and Shanghai Committee of Science and Technology (No.08411951200, No.08JG05002).

References

1. Aharon, M., Elad, M., Bruckstein, A.M.: The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54, 4311–4322 (2006)
2. Torralba, A., Fergus, R., Freeman, W.T.: Tiny Images. Technical Report, Computer Science and Artificial Intelligence Lab., MIT (2007)
3. Fellbaum, C.: Wordnet: An Electronic Lexical Database. Bradford Books (1998)
4. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing 15, 3736–3745 (2006)
5. Donoho, D.L., Elad, M., Temlyakov, V.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on Information Theory 147, 185–195 (2007)
6. Bryt, O., Elad, M.: Compression of facial images using the K-SVD algorithm. J. Vis. Commun. Image Represent 19(4), 270–283 (2008)
7. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. CVPR (June 2005)
8. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. ICCV (2005)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
10. Li, L., Wang, G., Fei-Fei, L.: Optimol: automatic online picture collection via incremental model learning. In: Proc. CVPR (June 2007)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
12. Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. BMVC, vol. 1, pp. 384–393 (2002)
13. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV 60(1), 63–86 (2004)
14. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transaction on PAMI 27(10), 1615–1630 (2005)
15. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proc. CVPR (June 2006)
16. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Proc. ICCV, pp. 370–377 (2005)
17. Yeh, T., Lee, J., Darrell, T.: Adaptive vocabulary forests for dynamic indexing and category learning. In: Proc. ICCV (October 2007)
18. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
19. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 43, 177–196 (2001)
20. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. SIGIR (1999)

21. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proc. CVPR (2009)
22. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: A database and web-based tool for image annotation. IJCV 77(1-3), 157–173 (2008)
23. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Conf. Rec. 27th Asilomar Conf. Signals, Syst. Comput., vol. 1 (1993)
24. Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theory 50, 2231–2242 (2004)
25. Donoho, D.: For Most Large Underdetermined Systems of Linear Equations the Minimal l1-Norm Solution Is Also the Sparsest Solution. Comm. Pure and Applied Math. 59(6), 797–829 (2006)
26. Candes, E., Romberg, J., Tao, T.: Stable Signal Recovery from Incomplete and Inaccurate Measurements. Comm. Pure and Applied Math. 59(8), 1207–1223 (2006)
27. Chen, S., Donoho, D., Saunders, M.: Atomic Decomposition by Basis Pursuit. SIAM Rev. 43(1), 129–159 (2001)

Semi-supervised Nearest Neighbor Discriminant Analysis Using Local Mean for Face Recognition

Caikou Chen[1], Pu Huang[1], and Jingyu Yang[2]

[1] Information Engineering College, Yangzhou University, Yangzhou 225009, China
[2] School of Computer Science and Technology Nanjing University of Science and Technology Nanjing, 210094, China

Abstract. Feature extraction is the key problem of face recognition. In this paper, we propose a new feature extraction method, called semi-supervised local mean-based discriminant analysis (SLMNND). SLMNND aims to find a set of projection vectors which respect the discriminant structure inferred from the labeled data points, as well as the intrinsic geometrical structure inferred from both labeled and unlabeled data points. Experiments on the famous ORL and AR face image databases demonstrate the effectiveness of our method.

1 Introduction

Face recognition is an important topic in the fields of pattern recognition and computer vision. Feature extraction (or dimensionality reduction) is the key problem in face recognition. Over the past decades, there have been lots of feature extraction techniques, among which, principal component analysis (PCA) [1] and linear fisher analysis (LDA) [2] are the two most popular ones.

PCA is an unsupervised method, which aims to find a set of orthogonal projection axes in the sense of minimum mean square error, but does not consider the class information and may thus be not reliable for classification task [3, 4]. Unlike PCA, LDA is a supervised method, which searches for a series of projection vectors that maximizes the inter-class scatter while minimizes the intra-class scatter. In general, when the label information is known, e.g. for classification task, LDA can achieve more significant performance than PCA [3, 4]. However, LDA is a linear algorithm, when the given data are nonlinearly distributed, LDA may fail to discover the essential intrinsic structure of the data [8, 9]. Recently, several manifold-learning based learning methods, with emphasis on local structure of the data, represented by LLE [5], LPP [6] and Local-Mean based Nearest Neighbor Discriminant (LMNND) [7] have attracted more and more attention on pattern recognition. These methods have shown their impressive performance both for data visualization and classification. According to the methods referred above, we can find that for classification problems, both local information and label inform are important.

In real world, the label information of the data are usually known little or unknown so that the classification ability of the methods on testing samples

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 331–338, 2010.

can not be guaranteed. In this case, a possible solution to deal with insufficient labeled samples could be learning on both labeled and unlabeled samples (e.g. semi-supervised learning).

In this paper, we proposed a semi-supervised and robust version of LMNND (SLMNND). SLMNND aims to find a set of projection vectors which respect the discriminant structure inferred from the labeled data points, as well as the intrinsic geometrical structure inferred from both labeled and unlabeled data points. Specifically, the labeled data points, combined with the unlabeled data points, are used to build a graph incorporating neighborhood information of the data set. The graph provides a discrete approximation to the local geometry of the data manifold. Using the notion of graph Laplacian, a smoothness penalty on the graph can be incorporated into the objective function. In this way our SLMNND algorithm can optimally preserves the manifold structure. Experiments on the well-known face image databases such as the ORL and AR databases validate the effectiveness of our proposed method.

2 Review of LMNND

Given a set of M samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M \in \mathbb{R}^n$, belonging to C classes, and let $\boldsymbol{x}_i^r$ denote the ith sample in class r. Suppose the K-nearest neighbors of $\boldsymbol{x}_i^r$ in class p are $\boldsymbol{x}_{ik}^p$, where $k = 1, 2, \ldots, K$. Let us calculate the mean vector of these K-nearest neighbors of $\boldsymbol{x}_i^r$ in class p as follows:

$$\boldsymbol{m}_p^{ir} = \sum_{k=1}^{K} \boldsymbol{x}_{ik}^p \tag{1}$$

The within-class local scatter matrix in LMNND is defined as:

$$\boldsymbol{S}_{\mathrm{w}}^{\mathbf{L}} = \sum_{r,i} (\boldsymbol{x}_i^r - \boldsymbol{m}_r^{ir})(\boldsymbol{x}_i^r - \boldsymbol{m}_r^{ir})^\top \tag{2}$$

Similarly, the between-class local scatter matrix is calculated by:

$$\boldsymbol{S}_{\mathrm{b}}^{\mathbf{L}} = \sum_{r,i} (\boldsymbol{x}_i^r - \boldsymbol{m}_l^{ir})(\boldsymbol{x}_i^r - \boldsymbol{m}_l^{ir})^\top \tag{3}$$

where $\boldsymbol{m}_l^{ir}$ is the minimal local mean of $\boldsymbol{x}_i^r$ in class $l (l \neq r)$ and the Euclidean distance between $\boldsymbol{x}_i^r$ and $\boldsymbol{m}_l^{ir}$ meets the condition:

$$d(\boldsymbol{x}_i^r, \boldsymbol{m}_l^{ir}) = \min_{p, p \neq r} d(\boldsymbol{x}_i^r, \boldsymbol{m}_p^{ir}) \tag{4}$$

The objective function of LMNND is defined as follows:

$$L_{LM-NND}(\varphi) = \underset{\varphi^\top \varphi}{\operatorname{argmax}} \frac{\varphi^\top \boldsymbol{S}_b^L \varphi}{\varphi^\top \boldsymbol{S}_w^L \varphi} \tag{5}$$

3 Semi-supervised Local-Mean Based Nearest Neighbor Discriminant Analysis (SLMNND)

3.1 Modification of LMNND

Suppose $\boldsymbol{x}_i$ is the ith column (sample) in the training set $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M]$, and ω_i denotes the class label of $\boldsymbol{x}_i$. After the projection of each sample onto the projection axis $\boldsymbol{\varphi}$, where $y_i = \boldsymbol{\varphi}^\top \boldsymbol{x_i}$, we will get the transformed images $\boldsymbol{Y} = [y_1, y_2, \ldots, y_M]$. If $\omega_i = \omega_j = p$ and $\boldsymbol{x}_j$ is one of the K-nearest neighbors of $\boldsymbol{x}_i$ in class p, the within-class local scatter of $\boldsymbol{x}_i$ in the transformed space is defined as $z_i = y_i - \frac{1}{K}\sum_{j=1}^{K} y_j$. The total within-class scatter can be calculated as:

$$f(y) = \sum_{i=1}^{M} (z_i)^2 \tag{6}$$

where

$$z_i = y_i - \sum_{j=1}^{K} \boldsymbol{W}_{ij} y_j \tag{7}$$

and

$$\boldsymbol{W}_{ij} = \begin{cases} \dfrac{1}{K} & \text{if } \omega_i = \omega_j = p \\ & \text{and } \boldsymbol{x}_j \text{ is one of the } K-\text{nearest} \\ & \text{neighbors of } \boldsymbol{x}_i \text{ in class } p \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Following some algebraic transformations, the total within-class scatter $f(y)$ can be converted into:

$$\begin{aligned} f(y) &= \sum_{i=1}^{M} (y_i - \tfrac{1}{K}\sum_{j=1}^{K} y_j)^2 = \sum_{i=1}^{M} z_i^2 = \boldsymbol{z}^\top \boldsymbol{z} \\ &= \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{W})^\top (\boldsymbol{I} - \boldsymbol{W}) \boldsymbol{y} \\ &= \boldsymbol{\varphi}^\top \boldsymbol{X} (\boldsymbol{I} - \boldsymbol{W})^\top (\boldsymbol{I} - \boldsymbol{W}) \boldsymbol{X}^\top \boldsymbol{\varphi} \\ &= \boldsymbol{\varphi}^\top \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^\top \boldsymbol{\varphi} \\ &= \boldsymbol{\varphi}^\top \boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} \boldsymbol{\varphi} \end{aligned} \tag{9}$$

where,

$$\boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} = \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^\top \tag{10}$$

$$\boldsymbol{M}_{\mathrm{w}} = (\boldsymbol{I} - \boldsymbol{W})^\top (\boldsymbol{I} - \boldsymbol{W}) \tag{11}$$

Similarly, for a given sample $\boldsymbol{x}_i$, its between-class local mean vectors are represented as $\{\boldsymbol{m}_i^r | i = 1, 2, \ldots, M, r = 1, 2, \ldots, C, r \neq \omega_i\}$, where $\boldsymbol{m}_i^r$ is the between-class local mean vector of $\boldsymbol{x}_i$ in class r. We can choose K_c nearest between-class local mean vectors of $\boldsymbol{x}_i$ from $\{\boldsymbol{m}_i^r\}$, and these chosen vectors are represented as

$\{\boldsymbol{m}_{iq}|i = 1, 2, \ldots, M, q = 1, 2, \ldots, K_c\}$, where $\boldsymbol{m}_{iq}$ represents the qth minimal between-class local mean vector of $\boldsymbol{x}_i$.

We define the elements of $\boldsymbol{B}^q$ by:

$$\boldsymbol{B}_{ij}^q = \begin{cases} \dfrac{1}{K} & \text{if } \omega_i = r \neq \omega_j,\ \boldsymbol{m}_i^p = \boldsymbol{m}_{iq}, \text{ and} \\ & \boldsymbol{x}_j \ \text{ is one of the } K-\text{nearest} \\ & \text{neighbors of } \boldsymbol{x}_i \text{ in class } p. \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Thus, the modified between-class scatter in the transformed space can be rewritten as:

$$\begin{aligned} \tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} &= \sum_{q=1}^{K_c} \sum_{i=1}^{M} (\boldsymbol{x}_i - \boldsymbol{m}_{iq})(\boldsymbol{x}_i - \boldsymbol{m}_{iq})^\top \\ &= \sum_{q=1}^{K_c} \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{B}^q)^\top (\boldsymbol{I} - \boldsymbol{B}^q)\boldsymbol{X}^\top \\ &= \boldsymbol{X}\boldsymbol{M}_{\mathrm{b}}\boldsymbol{X}^\top \end{aligned} \tag{13}$$

where

$$\boldsymbol{M}_{\mathrm{b}} = \sum_{q=1}^{K_c} (\boldsymbol{I} - \boldsymbol{B}^q)^\top (\boldsymbol{I} - \boldsymbol{B}^q) \tag{14}$$

The modified LMNND discriminant criterion function can be defined by:

$$J_{NLM-NND}(\boldsymbol{\varphi}) = \underset{\boldsymbol{\varphi}^\top\boldsymbol{\varphi}}{\operatorname{argmax}} \frac{\boldsymbol{\varphi}^\top \tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} \boldsymbol{\varphi}}{\boldsymbol{\varphi}^\top \boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} \boldsymbol{\varphi}} \tag{15}$$

3.2 Semi-supervosed Extension of Modified LMNND

The modified LMNND aims to find a projection vector $\boldsymbol{\varphi}$ such that the ratio of $\boldsymbol{\varphi}^\top \tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} \boldsymbol{\varphi}$ and $\boldsymbol{\varphi}^\top \boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} \boldsymbol{\varphi}$ is maximized. When there is no sufficient training sample, overfitting may happen. A typical way to prevent overfitting is to impose a regularizer. The optimization problem of the regularized version of the modified LMNND can be written as follows:

$$\underset{\boldsymbol{\varphi}^\top\boldsymbol{\varphi}}{\operatorname{argmax}} \frac{\boldsymbol{\varphi}^\top \tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} \boldsymbol{\varphi}}{\boldsymbol{\varphi}^\top \boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} \boldsymbol{\varphi} + \eta J(\boldsymbol{\varphi})} \tag{16}$$

where $J(\boldsymbol{\varphi})$ controls the learning complexity of the hypothesis family, and the coefficient η controls a balance between the model complexity and the empirical loss.

Given a set of examples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M]$, we can use a p-nearest neighbor graph G to model the relationship between nearby data points. Specifically, we put an edge between nodes i and j if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are "close", i.e., $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are

among p nearest neighbors of each other. Let the corresponding weight matrix be $\boldsymbol{H}$, defined by:

$$\boldsymbol{H}_{ij} = \begin{cases} 1 & \textit{if } \boldsymbol{x}_i \in N_p(\boldsymbol{x}_j) \textit{ or } \boldsymbol{x}_j \in N_p(\boldsymbol{x}_i) \\ 0 & \textit{otherwise} \end{cases} \tag{17}$$

where $N_p(\boldsymbol{x}_j)$ denotes the p-nearest neighbors of $\boldsymbol{x}_i$. In general, the mapping function should be as smooth as possible on the graph. Specifically, if two data points are linked by an edge, they are likely to be in the same class. Moreover, the data points lying on a densely linked sub-graph are likely to have the same label. Thus, a natural regularizer can be defined as follows:

$$\begin{aligned} J(\boldsymbol{\varphi}) &= \tfrac{1}{2}\sum_{ij} \boldsymbol{H}_{ij}(\boldsymbol{\varphi}^\top \boldsymbol{x}_i - \boldsymbol{\varphi}^\top \boldsymbol{x}_j)^2 \\ &= \sum_i \boldsymbol{\varphi}^\top \boldsymbol{x}_i \boldsymbol{D}_{ij} \boldsymbol{x}_i^\top \boldsymbol{\varphi} - \sum_{ij} \boldsymbol{\varphi}^\top \boldsymbol{x}_i \boldsymbol{H}_{ij} \boldsymbol{x}_j^\top \boldsymbol{\varphi} \\ &= \boldsymbol{\varphi}^\top \boldsymbol{X}(\boldsymbol{D} - \boldsymbol{H})\boldsymbol{X}^\top \boldsymbol{\varphi} \\ &= \boldsymbol{\varphi}^\top \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top \boldsymbol{\varphi} \end{aligned} \tag{18}$$

where $\boldsymbol{D}$ is a diagonal matrix,and each diagonal entry is the column (or row, since $\boldsymbol{H}$ is symmetric) sum of $\boldsymbol{H}$, namely $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{H}_{ij}$, and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{H}$ is called the Laplacian matrix.

With this data dependent regularizer, the generalized objective function of SLMNND can be defined by:

$$J_{SLMNND}(\boldsymbol{\varphi}) = \underset{\boldsymbol{\varphi}^\top \boldsymbol{\varphi}}{\operatorname{argmax}} \frac{\boldsymbol{\varphi}^\top \tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} \boldsymbol{\varphi}}{\boldsymbol{\varphi}^\top (\boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} + \eta \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top)\boldsymbol{\varphi}} \tag{19}$$

The projective vector $\boldsymbol{\varphi}$ that maximizes the objective function is given by the maximum eigenvalue solution of the generalized eigenvalue problem:

$$\tilde{\boldsymbol{S}}_{\mathrm{b}}^{\mathrm{L}} \boldsymbol{\varphi} = \lambda(\boldsymbol{S}_{\mathrm{w}}^{\mathrm{L}} + \eta \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top)\boldsymbol{\varphi} \tag{20}$$

3.3 The Algorithm

Given a labeled set $\{\boldsymbol{x}_i\}_{i=1}^{l}$ belonging to C classes and an unlabeled set $\{\boldsymbol{x}_i\}_{i=l+1}^{M}$. The algorithm procedure of the proposed methods is stated below:

1. **Construct the adjacency graph:** Construct the p-nearest neighbors graph matrix $\mathbf{H}$ in Eq. (18) and calculate the graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{H}$.
2. **Construct the labeled graph:** Construct the weight matrix $\mathbf{W}$ and $\boldsymbol{B}^q$ using Eq. (8) and Eq. (12), and then we can get $\boldsymbol{M}_w$ and $\boldsymbol{M}_b$ using Eq. (11) and Eq. (14) respectively. Then the labeled graph can be constructed as:

$$\tilde{\boldsymbol{M}}_b = \begin{pmatrix} \boldsymbol{M}_b & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{\boldsymbol{M}}_w = \begin{pmatrix} \boldsymbol{M}_w & 0 \\ 0 & 0 \end{pmatrix}$$

3. **Eigen-problem:** Compute the eigenvectors with respect to the non-zero eigenvalues for the generalized eigenvector problem:

$$\boldsymbol{X}\tilde{\boldsymbol{M}}_b\boldsymbol{X}^\top\boldsymbol{\varphi} = \lambda\boldsymbol{X}(\tilde{\boldsymbol{M}}_w + \eta\boldsymbol{L})\boldsymbol{X}^\top\boldsymbol{\varphi} \tag{21}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_M]$.

4. **SLMNND Embedding:** Let $\Phi = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_d]$,an $n \times d$ transformation matrix and $\boldsymbol{\varphi}_i (i = 1, \ldots, d)$ is a vector corresponding to the ith largest non-zero eigenvalue computed by Step 3. The samples can be embedded into d dimensional subspace by $\boldsymbol{Y} = \Phi^\top\boldsymbol{X}$.

4 Experimental Results

In this section, we test the effectiveness of the proposed SLMNND method on face recognition problem. We compare SLMNND with PCA, LDA, LPP and LMNND on the ORL and AR face database. Since the dimensions of the face images are large, we apply PCA to reduce their dimensionality and remove the noise in these databases before we perform these algorithms.

4.1 Experiments on the ORL Database

The ORL face database contains images from 40 individuals, each providing 10 different images. For some subjects, the images are taken at different times. The facial expressions (glasses or no glasses) also vary. The images are taken with a tolerance for tilting and rotation of the face of up to 20^o. Moreover, there is also some variation in the scale of up to about 10%. All images are grayscale and normalized to a resolution of 112×92 pixels.

In this experiment, we take the first 5 images each person as the training set, the rest 4 images as the testing set. Among these 5 images, $q = 1, 2, 3$ images are labeled and the rest $5 - q$ images are unlabeled respectively. First we choose the dimensions of PCA subspace as 30, 35, 40 corresponding to different training sets, then the parameters in our proposed method are set as $K = 1, K_c = 40/2 = 20$ and $p = 5 - 1 = 5, \eta = 1$. Table 1 shows the optimal recognition rates of the 5 methods corresponding to different training samples per class, and the baseline approach is the nearest neighbor classifier on the original image space. From the table 1, we can see that the proposed methods outperforms other four methods in the recognition rate cases. When $q = 1$, LDA and LMNND fail to work due to the singularity of their intra-class scatter matrices.

4.2 Experiments on the AR Database

The AR face database contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusion. The pictures of 120 individuals (65 men and 55 women) were taken in two sessions (separated by two weeks) and each section contains 13 color images. In our experiment, each image was manually cropped and resized to 50 40 pixels.

Table 1. Optimal recognition rates and the corresponding dimensions of the methods with different training samples per class

Method	Labeled training samples per class		
	1	2	3
Baseline	69.00	80.00	83.50
PCA	69.50(40)	80.00(40)	83.50
LDA	NULL	81.50(18)	85.50(19)
LPP	70.50(25)	81.50(29)	83.50(29)
LMNND	NULL	80.50(32)	85.50(20)
Proposed method	70.50(30)	86.00(31)	86.50(29)

The proposed method is tested on a subset of the AR database. This subset includes 1200 images of 120 individuals (the first 10 images of each person are chosen from the total AR database). In the experiments, we take the first 5 images per class as the training set, the rest 5 images as the testing set. Among the 5 training images per class, $q = 2, 3$ images are labeled and the rest $5 - q$ images are unlabeled. First the dimensions of PCA subspace are both 120 for LDA, LPP, LMNND and SLMNND in two cases. In our proposed method, the parameters are set a $Kc = 120/2 = 60$ and $p = 5 - 1 = 4, \eta = 1$, and for simplicity, $K = 1$.

Table 2. Optimal recognition rates and the corresponding dimensions of the methods with different training samples per class

Method	Labeled training samples per class	
	2	3
Baseline	80.00	78.67
PCA	79.67	77.50
LDA	87.83	93.67
LPP	93.50	94.50
LMNND	88.33	91.17
Proposed method	94.17	95.67

From table 1, we can see that the proposed method achieves the best recognition rate in all approaches.

5 Conclusions

In this paper, we proposed a novel semi-supervised learning-based feature extraction algorithm named Semi-supervised Local-Mean Based Nearest Neighbor Discriminant Analysis (SLMNND). SLMNND can efficient take use of the labeled

and unlabeled data points. The experimental results show that the proposed method is more effective and robust than the other methods.

Acknowledgments

This research was supported by the National Science Foundation of China, under Grant No. 60875004.

References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition. J. Cognitive Neuroscience 4(1), 71–86 (1991)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs fisherfaces: recognition using class specific linear projection. IEEE Trans. on PAMI 19(7), 711–720 (1997)
3. Liu, X., Wang, Z., Feng, Z.: Average neighborhood margin maximization projection with smooth regularization for face recognition. In: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, pp. 12–15 (July 2008)
4. Cai, D., He, X., Han, J.: Semi-Supervised discriminant analysis. In: 11th International Conference on IEEE, October 14-21, pp. 1–7 (2007)
5. Reweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
6. He, X., Yan, S., Hu, Y., NIyogi, P., Zhang, H.: Face recognition using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
7. Yang, J., Yang, J., Jin, Z.: New Concept for Discriminator Design: From Classifier to Discriminator. Pattern Recognition 22(24), 1–6 (2008)
8. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Grahph Embedding and Extension: Ageneral Framework for Dimensionality Reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1), 40–50 (2007)
9. Yang, J., Zhang, D.: Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(4), 650–664 (2007)

Decentralized Cohesive Motion Control of Multi-agent Formation in 3-Dimensional Space

Ran Zhao, Yongguang Yu, and Guoguang Wen

Department of Mathematics, Beijing Jiaotong University,
Beijing, China
{RanZhao,YongguangYu,GuoguangWen,lncs}@springer.com
http://www.springer.com/lncs

Abstract. In this paper, we generalize a set of decentralized control laws for cohesive motion of 3-dimensional multi-agent formation based on point-agent system model, which has been originally introduced for 2-dimensional formation. According to the numbers of degree of freedom, we investigate the persistency of multi-agent formation system. Then we design the corresponding control laws for the different class agents in the multi-agent formation. As a result, the 3-dimensional persistent formation can move cohesively along the trajectory as the target of the formation moves continuously along the prescribed trajectory. The effectiveness of the proposed control laws is demonstrated via simulation results.

Keywords: multi-agent formation; decentralized cohesive motion; point-agent; graph theory; 3-dimensional space.

1 Introduction

In recent years, the problem of multi-agent formation has attracted considerable attention as witnessed by explosion of papers in the area. Multi-agent formation system is widely applied in a variety of fields. Taking something in detail, in applications like exploration, surveillance and tracking, the multi-agent formations can be considered as a group of combat robots, teams of manned or unmanned aerial, space or underwater vehicles, etc. (1, 2, 3, 4, 5, 6, 7).

In the recent literatures, most of the works about the multi-agent formation system mainly focus on the following researches: the formation structure, including acquisition, maintenance and interconnected systems' stability of formation (5, 6, 7, 8, 9, 10, 11), the formation operation, including formation merging, formation splitting, and closing ranks (12, 13, 14, 15, 16), the formation coordination and control, including types of agent dynamics, flocking, cohesive motion, obstacle avoidance and so on(17, 18, 19, 20, 21, 22, 23, 24).

The formation coordination and control are particularly active areas of multi-agent formation system since many cooperative tasks are performed by controlling the agents to form up and move in a specified pattern. In (24), a theoretical framework that consists of graph theory and Lyapunov-based approaches to stability analysis and distributed control of multi-agent formations is provided, in

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 339–348, 2010.

which the formation control of the whole system is implemented via controlling each agent respectively.

In a slightly different context, in (19), a set of decentralized control laws for the cohesive motion of 2-dimensional multi-agent formation have been presented(the aim is to move a given persistent formation with specified initial position and orientation to a new desired position and orientation cohesively, i.e., without violating the persistence of the formation during the motion.). Based on (19), two strategies for obstacle avoidance by the entire formation have been further studied in (20).

Most of the above work are based on the 2-dimensional space. However, in the real world, the applications of multi-agent formation are often in 3-dimensional space as opposed to the plane. Furthermore, in 3-dimensional space, there are many characteristics which differ from that in 2-dimensional space. The multi-agent formation problem is much more complex in 3-dimensional space. Thus, we focus on the decentralized control laws for the cohesive motion of 3-dimensional multi-agent formation in this paper, where each of the agents makes decisions based only on its own observations and state. Considering a point-agent system model, we categorize the agents according to the numbers of degrees of freedom(DOFs) of them, *i.e.*, we label the agents as 0-DOF agents, 1-DOF agents, 2-DOF agents, 3-DOF agents, respectively. Then we design the corresponding control laws to the different class agents, distinctively. Furthermore, we let the target of the formation move continuously along the prescribed trajectory, and the 3-dimensional persistent formation can move cohesively along the trajectory. The effectiveness of the proposed control laws is illustrated by some simulations.

The structure of this paper is organized as follows: the section 2 provides the definition of the problem; in section 3, the proposed control framework is discussed; in section 4, some results of simulations are presented. The paper ends with concluding remarks in section 5.

2 Problem Definition

In this section, we introduce some important notions, and also give some assumptions relevant to our work. We mainly consider the 3-dimensional point-agent persistent formation.

The formation graph of multi-agent is a directed graph $G = (V, E)$, which consists of a set of vertices V and a directed edge set E. It can be used to depict the control architecture as follows: each agent corresponds a vertex in V, and for each agent(vertex) pair (A_i, A_j), there is a directed edge $\overrightarrow{(A_i, A_j)}$ from A_i to A_j if A_i has a constraint on the distance it must actively maintain from A_j, in which case only can we say that A_i follows A_j (or A_i is a follower of A_j).

A formation is named rigid if it, under the continuous motion, satisfies that the distance of each agent pairs explicitly maintains constant, then the shape of the formation will be maintained. A rigid formation is further called minimally rigid if no single inter-agent distance constraint can be removed without losing rigidity. The notions we refer above are all defined in the undirected graph.

Correspondingly, in the directed graph, the persistence is defined, corresponding to the rigidity in the undirected graph. A formation is called persistent if it is rigid and satisfies another condition named constraint consistence, which is equivalent to the requirement that it is possible to maintain the nominated inter-agent distances. Similarly, a formation is called minimally persistent if it is minimally rigid and constraint consistent. Rigorous definitions and detailed characteristics of rigidity, minimally rigidity, constraint consistence, persistence and minimally persistence can be found in Refs.([5], [6], [7], [8]).

Generally, we use the term out-degree of a vertex A_i to denote the number of distance constraints. Associated with the out-degree, a rather significant notion, the number of degrees of freedom of a vertex A_i refers to the dimension of the space in which the corresponding agent can choose its position. Usually, we abbreviate degree of freedom as DOF. The number of DOFs of a vertex A_i in R^3 is given by $max(0, 3 - d^+(A_i))$, where $d^+(A_i)$ represents the out-degree of the vertex A_i. For example, in R^3, the numbers of DOFs of the vertices with zero, one, and two out-degrees are 3, 2, and 1, respectively; all the other vertices have 0 DOF.

In 2-dimensional space, for a persistent formation, it has been shown in ([7], [19]),that the sum of DOFs of all the individual agents is at most 3 and for a minimally persistent formation, exactly 3. Based on the distribution of these 3 DOFs, minimally persistent formations can be divided into two categories: $S_1 = \{2, 1, 0, 0,...\}$, $S_2 = \{1, 1, 1, 0, 0,...\}$. S_1 is called formation with the leader-follower structure where one agent has 2 DOFs, another has 1 DOF and the rest have 0 DOF. S_2 is called formation with the 3-coleader structure where three agents have 1 DOF respectively and the rest have 0 DOF. Similarly, in 3-dimensional space, the sum of DOFs of all the individual agents is at most 6, and for a minimally persistent formation,it is exactly 6. Based on the distribution of these 6 DOFs, minimally persistent formations can be divided into the following categories: $S_1 = \{3, 2, 1, 0, 0,...\}$, $S_2 = \{2, 2, 2, 0, 0,...\}$, $S_3 = \{3, 1, 1, 1, 0, 0,...\}$, $S_4 = \{2, 2, 1, 1, 0, 0,...\}$, $S_5 = \{2, 1, 1, 1, 1, 0, 0,...\}$, $S_6 = \{1, 1, 1, 1, 1, 1, 0, 0,...\}$. In the above categories, a last possible category is skipped, $S_0 = \{3, 3, 0, 0,...\}$, since it is an undesired one that can be accepted only temporarily during formation changes as a transient state. This is because S_0 apparently allows two leaders simultaneously in a formation, corresponding to failure of structural persistence and we want the DOFs assignment to avoid this state. Further detailed interpretations can be found in Refs.([6], [7]).

As mentioned above, providing that the numbers of DOFs of the agents are different, the numbers of the constraints of them are also different. If some agent A_i has 3 DOFs then it can move freely in R^3, If the agent A_i has 2 DOFs, then it can move freely on the spherical surface with center A_j and radius d_{ij}, where A_j is the agent which A_i follows and d_{ij} is the distance between them. If the agent A_i has 1 DOF, then it can rotate freely on the circle, which is the intersection of two spherical surfaces with A_j, A_k as the centers and d_{ij}, d_{ik} as the radiuses, respectively. If the agent A_i has 0 DOF, then it can move depending on three agents it follows completely.

In the paper, we mainly consider a formation F with agents $A_i, \ldots, A_m$ ($m \geq 3$). Assume each of the agents $A_i, \ldots, A_m$ are point agents whose positions on the 3-dimentional (x, y, z) Cartesian coordinate system at time t are given by $p_i(t) = (x_i(t), y_i(t), z_i(t)) \in R^3$, $i \in \{1, \ldots, m\}$. The control task is to move a given persistent formation F, whose initial position and orientation are specified, to a new desired position and orientation in R^3 cohesively, *i.e.* without violating the persistence of the formation F during its motion,using a decentralized strategy. The initial and final positions and orientations of F are defined by the initial positions p_{i0} and final positions p_{if} of the individual agents A_i for $i \in \{1, \ldots, m\}$, respectively. Assume that the final positions p_{if} are consistent with the desired inter-agent distances d_{ij} between neighbor agent pairs (A_i, A_j).

In a given formation, assume that each agent A_i satisfies the following velocity integrator kinematics:

$$\dot{p}_i(t) = v_i(t) \tag{1}$$

where $v_i(t) = (v_{xi}(t), v_{yi}(t), v_{zi}(t)) \in R^3$ denotes velocity of A_i at time t. The velocity $v_i(t)$ is considered as the control signal to be generated by the individual controller of agent A_i. It is required that $v_i(t)$ is continuous and satisfies $\mid v_i(t) \mid < \bar{v}$, for some constant maximum speed limit $\bar{v} > 0$ at any $t \geq 0$ for any $i \in \{1, \ldots, m\}$.

According to the requirement for the task of control, we give several assumptions as below:

1. Each agent A_i in the formation can know its own position $p_i(t)$, velocity $v_i(t)$, final desired position p_{if} as well as the position $p_j(t)$ of any agent A_j that it follows at any time $t \geq 0$. In addition, it is also assumed that the distance sensing range for a neighbor agent pair (A_i, A_j) is sufficiently larger than the desired distance d_{ij} to be maintained.

2. The motion of every agent in the formation at any time just relies on the agents it follows and the final position it desires to reach.

3. For each 1-DOF agent or each 2-DOFs agent, we design the control laws as following such that meeting the distance constraints has a higher priority to reaching to the final desired position, i.e., 1-DOF agents or 2-DOFs agents can undergo their DOFs movements, only when their distance constraints are satisfied within a certain error bound.

Based on the agent kinematics (1) and the assumptions above, in the next section, we develop control laws for the cohesive motion of 3-dimensional multi-agent formation.

3 Control Law for Agents with Different DOFs

According to the previous section, we know that providing the numbers of DOFs of the agents differ, the numbers of the constraints of them are distinct. Furthermore, we can classify the agents in terms of the numbers of DOFs of them, that is, we can label the agents as 0-DOF agents, 1-DOF agents, 2-DOF agents, 3-DOF agents, respectively.

In this section, for the cohesive motion of a persistent formation, we design various control laws to the above classes, distinctively.

3.1 Control Law for the 0-DOF Agents

Consider a 0-DOF agent A_i and the three agents A_j, A_k and A_l it follows. Due to the distance constraints of keeping $\mid p_i(t)-p_j(t) \mid$, $\mid p_i(t)-p_k(t) \mid$, $\mid p_i(t)-p_l(t) \mid$ at the desired values d_{ij}, d_{ik}, d_{il} respectively, at any time $t \geq 0$, the desired position $p_{id}(t)$ of A_i is the point whose distances to $p_j(t)$, $p_k(t)$, $p_l(t)$ are d_{ij}, d_{ik}, d_{il} respectively. Further, $p_{id}(t)$ must vary continuously. Assuming $\mid p_i(t) - p_{id}(t) \mid$ is sufficiently small, $p_{id}(t)$ can be explicitly determined as $p_{id}(t) = \bar{p}_{jkl}(t, p_i(t))$, where $\bar{p}_{jkl}(t, p_0)$ for any $p_0 \in R^3$ denotes the intersection of the spherical surface $O(p_j(t), d_{ij})$, $O(p_k(t), d_{ik})$, and $O(p_l(t), d_{il})$ that is closer to p_0, and in the notion $O(\cdot, \cdot)$ the first argument denotes the center and the second denotes the radius. Based on this observation, we use the following control law for the 0-DOF agent A_i :

$$v_i(t) = \bar{v}\beta_i(t)\delta_{id}(t)/|\delta_{id}(t)| \tag{2}$$

$$\delta_{id}(t) = p_{id}(t) - p_i(t) = \bar{p}_{jk}(t, p_i(t)) - p_i(t)$$

$$\beta_i(t) = \begin{cases} 0, & |\delta_{id}(t)| < \varepsilon_j \\ \frac{|\delta_{id}(t)|-\varepsilon_j}{\varepsilon_j} & \varepsilon_j \leq |\delta_{id}(t)| < 2\varepsilon_j \\ 1, & |\delta_{id}(t)| \geq 2\varepsilon_j \end{cases}$$

where $\bar{v} > 0$ is the constant maximum speed of the agents, and $\varepsilon_j > 0$ is a small design constant. In (2), the switching term $\beta_i(t)$ is introduced to avoid chattering due to small but acceptable errors in the desired inter-agent distances.

3.2 Control Law for the 1-DOF Agents

Consider a 1-DOF agent A_i and the two agents A_j, A_k it follows. A_i requires to maintain its distances to A_j, A_k. It is free to move on the circle which is the intersection of two spherical surfaces with the centers A_j, A_k, the radiuses d_{ij}, d_{ik}, respectively. It does not need to use the whole of its velocity capacity to satisfy its distance constraints. Considering that the agent should satisfy two constraints of distance, without losing generality, we assume that the two constraints are in order, namely one after another. Based on this observation and the assumption, we propose the following control scheme for the 1-DOF agent A_i :

$$v_i(t) = \beta_i(t)v_{i1}(t) + \sqrt{1-\beta_i^2(t)}\{\bar{\beta}_i(t)v_{i2}(t)$$

$$+\sqrt{1-\bar{\beta}_i^2(t)}(\bar{\bar{\beta}}_i(t)v_{i3}(t))\}$$

$$v_{i1}(t) = \bar{v}sgn(\bar{\delta}_{ji}(t))\delta_{ji}(t)/|\delta_{ji}(t)| \tag{3}$$

$$v_{i2}(t) = \bar{v}sgn(\bar{\delta}_{ki}(t))\delta_{ki}(t)/|\delta_{ki}(t)| \tag{4}$$

$$v_{i3}(t) = \bar{v}\bar{\delta}^*_{ji}(t) \tag{5}$$

$$\bar{\delta}^*_{ji}(t) = (\delta_{if}(t) - \frac{\delta_{ji}(t)}{\mid \delta_{ji}(t) \mid} \cdot \mid \delta_{if}(t) \mid \cdot \cos\theta)/|(\delta_{if}(t)$$

$$-\frac{\delta_{ji}(t)}{\mid \delta_{ji}(t) \mid} \cdot \mid \delta_{if}(t) \mid \cdot \cos\theta)|$$

$$\cos\theta = (\delta_{if}(t)\delta_{ji}(t))/(|\delta_{if}(t)|\delta_{ji}(t)|)$$

$$\delta_{if}(t) = p_{if}(t) - p_i(t)$$

$$\delta_{ji}(t) = p_j(t) - p_i(t) = (\delta_{jix}(t), \delta_{jiy}(t), \delta_{jiz}(t))$$

$$\bar{\delta}_{ji}(t) = |\delta_{ji}(t)| - d_{ij}$$

$$\delta_{ki}(t) = p_k(t) - p_i(t) = (\delta_{kix}(t), \delta_{kiy}(t), \delta_{kiz}(t))$$

$$\bar{\delta}_{ki}(t) = |\delta_{ki}(t)| - d_{ik}$$

$$\beta_i(t) = \begin{cases} 0, & |\bar{\delta}_{ji}(t)| < \varepsilon_j \\ \frac{|\bar{\delta}_{ji}(t)|-\varepsilon_j}{\varepsilon_j}, & \varepsilon_j \le |\bar{\delta}_{ji}(t)| < 2\varepsilon_j \\ 1, & |\bar{\delta}_{ji}(t)| \ge 2\varepsilon_j \end{cases}$$

$$\bar{\beta}_i(t) = \begin{cases} 0, & |\bar{\delta}_{ki}(t)| < \varepsilon_k \\ \frac{|\bar{\delta}_{ki}(t)|-\varepsilon_k}{\varepsilon_k}, & \varepsilon_k \le |\bar{\delta}_{ki}(t)| < 2\varepsilon_k \\ 1, & |\bar{\delta}_{ki}(t)| \ge 2\varepsilon_k \end{cases}$$

$$\bar{\bar{\beta}}_i(t) = \begin{cases} 0, & |\delta_{if}(t)| < \varepsilon_f \\ \frac{|\delta_{if}(t)|-\varepsilon_f}{\varepsilon_f}, & \varepsilon_f \le |\delta_{if}(t)| < 2\varepsilon_f \\ 1, & |\delta_{if}(t)| \ge 2\varepsilon_f \end{cases}$$

ε_j, ε_k, $\varepsilon_f > 0$ are small design constants. the switching term $\beta_i(t)$, $\bar{\beta}_i(t)$, and $\bar{\bar{\beta}}_i(t)$ are introduced to switch translational action (3), (4) to satisfy $|A_iA_j| \cong d_{ij}$, $|A_iA_k| \cong d_{ik}$, and a rotational action (5). $\bar{\delta}^*_{ji}(t)$ is the unit vector perpendicular to the distance vector $\delta_{ji}(t) = p_j(t) - p_i(t)$ in the plane formed by vector $\delta_{ji}(t)$ and $\delta_{if}(t)$, and we denote the angle of $\delta_{ji}(t)$ and $\delta_{if}(t)$ by θ.

3.3 Control Law for the 2-DOF Agents

Consider a 2-DOF agent A_i which only maintains its distance to A_j. It is free to move on the spherical surface with the center A_j and the radius d_{ij} provided that it does not need to use the whole of its velocity capacity to satisfy $|A_iA_j| = d_{ij}$. Based on this observation, we propose the following control scheme for the 2-DOF agent A_i :

$$v_i(t) = \beta_i(t)v_{i1}(t) + \sqrt{1-\beta_i^2(t)}v_{i2}(t) \tag{6}$$

$$v_{i1}(t) = \bar{v}sgn(\bar{\delta}_{ji}(t))\delta_{ji}(t)/|\delta_{ji}(t)| \tag{7}$$

$$v_{i2}(t) = \bar{v}\bar{\beta}_i(t)\bar{\delta}^*_{ji}(t) \tag{8}$$

$$\delta_{ji}(t) = p_j(t) - p_i(t) = (\delta_{jix}(t), \delta_{jiy}(t), \delta_{jiz}(t))$$

$$\bar{\delta}_{ji}(t) = |\delta_{ji}(t)| - d_{ij}$$

$$\bar{\delta}^*_{ji}(t) = (\delta_{if}(t) - \frac{\delta_{ji}(t)}{\mid \delta_{ji}(t) \mid} \cdot \mid \delta_{if}(t) \mid \cdot \cos\theta)/|(\delta_{if}(t)$$

$$-\frac{\delta_{ji}(t)}{\mid \delta_{ji}(t) \mid} \cdot \mid \delta_{if}(t) \mid \cdot \cos\theta)|$$

$$\cos\theta = (\delta_{if}(t)\delta_{ji}(t))/(|\delta_{if}(t)|\delta_{ji}(t)|)$$

$$\delta_{if}(t) = p_{if}(t) - p_i(t)$$

$$\beta_i(t) = \begin{cases} 0, & |\bar{\delta}_{ji}(t)| < \varepsilon_j \\ \frac{|\bar{\delta}_{ji}(t)|-\varepsilon_j}{\varepsilon_j}, & \varepsilon_j \le |\bar{\delta}_{ji}(t)| < 2\varepsilon_j \\ 1, & |\bar{\delta}_{ji}(t)| \ge 2\varepsilon_j \end{cases}$$

$$\bar{\beta}_i(t) = \begin{cases} 0, & |\delta_{if}(t)| < \varepsilon_f \\ \frac{|\delta_{if}(t)|-\varepsilon_f}{\varepsilon_f}, & \varepsilon_f \le |\delta_{if}(t)| < 2\varepsilon_f \\ 1, & |\delta_{if}(t)| \ge 2\varepsilon_f \end{cases}$$

ε_j, $\varepsilon_f > 0$ are small design constants. In (6), via the switching term $\beta_i(t)$, the controller switches between a translational action (7) to satisfy $|A_iA_j| \cong d_{ij}$, and a rotational action (8) to move the agent A_i towards p_{if}, which can take place only when $|A_iA_j|$ is sufficiently close to d_{ij}.

In (8), $\bar{\delta}^*_{ji}(t)$ is the unit vector perpendicular to the distance vector $\delta_{ji}(t) = p_j(t) - p_i(t)$ in the plane formed by vector $\delta_{ji}(t)$ and $\delta_{if}(t)$, and we denote the angle of $\delta_{ji}(t)$ and $\delta_{if}(t)$ by θ. The switching term $\bar{\beta}_i(t)$ is for avoiding chattering due to small but acceptable errors in the final position of A_i.

3.4 Control Law for the 3-DOF Agents

For a 3-DOF agent A_i, since it does not have any constraints to satisfy, it can use its full velocity capacity only to move towards its desired final position p_{if}. Hence the velocity input at any time t can be simply designed as a vector with the constant maximum speed $\bar{v}$ in the direction of $p_{if}(t) - p_i$.

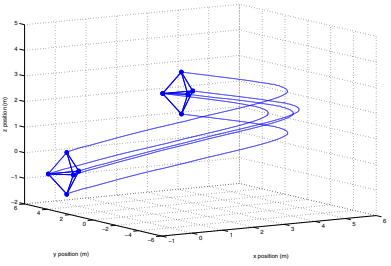
$$v_i(t) = \bar{v}\bar{\beta}_i(t)\delta_{if}(t)/|\delta_{if}(t)| \tag{9}$$

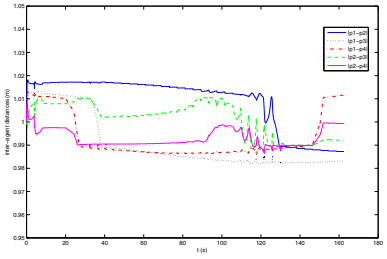
$$\delta_{if}(t) = p_{if}(t) - p_i(t)$$

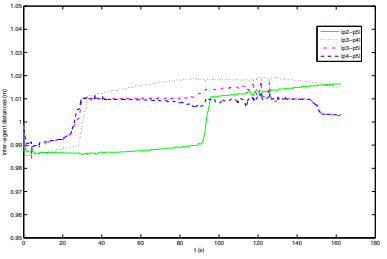
$$\bar{\beta}_i(t) = \begin{cases} 0, & |\delta_{if}(t)| < \varepsilon_f \\ \frac{|\delta_{if}(t)|-\varepsilon_f}{\varepsilon_f}, & \varepsilon_f \le |\delta_{if}(t)| < 2\varepsilon_f \\ 1, & |\delta_{if}(t)| \ge 2\varepsilon_f \end{cases}$$

The switching term $\bar{\beta}_i(t)$ again prevents chattering due to small but acceptable errors in the final position of A_i.

4 Simulations







5 Conclusion

In this paper, we have presented a set of decentralized control laws for the cohesive motion of 3-dimensional multi-agent formation based on point-agent system model, which has been originally introduced for 2-dimensional formation. We consider persistent formation that can be modeled by directed graphs. The effectiveness of the designed control schemes have been demonstrated via a set of simulation results.

As further extensions of this study, there exist a number of topics for future work. The control and sensing in the simulations are assumed to be noiseless in this paper. In real world applications, however, the scheme clearly has to be implemented to be robust to noise, and scalability and complexity are among other aspects to be considered. As to optimization of the current approach, the control scheme can be re-designed for agent kinematics other than the velocity integrator model. Since the operations of most of the multi-agent systems are naturally delayed, the study of the formation system with delay needs more attention. In addition, obstacle avoidance for the formation system is yet another future problem.

Acknowledgment

The work described in this paper is fully supported by the National Nature Science Foundation of China (Grant No. 10901014) and the Research Fund for the Doctoral Program of Higher Education (No. 20070004030).

References

1. Anderson, B.D.O., Dasgupta, S., Yu, C.: Control of directed formations with a leader-first follower structure. In: Proc. IEEE Conference on Decision and Control, pp. 2882–2887 (2007)
2. Anderson, B.D.O., Yu, C., Fidan, B., Hendrickx, J.M.: Control and Information Architectures for Formations. In: Proc. IEEE Conference on Control Applications, pp. 1127–1138 (2006)
3. Gazi, V., Fidan, B.: Coordination and Control of Multi-agent Dynamic Systems: Models and Approaches. In: Şahin, E., Spears, W.M., Winfield, A.F.T. (eds.) SAB 2006 Ws 2007. LNCS, vol. 4433, pp. 71–102. Springer, Heidelberg (2007)
4. Anderson, B.D.O., Yu, C., Fidan, B., Hendrickx, J.: Use of meta-formations for cooperative control. In: Proc. International Symposium on Mathematical Theory of Networks and Systems, pp. 2381–2387 (2006)
5. Eren, T., Anderson, B.D.O., Whiteley, W., Morse, A.S., Belhumeur, P.N.: Operations on rigid formations of autonomous agents. Communications in On formations and Systems 3, 223–258 (2004)
6. Fidan, B., Yu, C., Anderson, B.D.O.: Acquiring and maintaining persistence of autonomous multi-vehicle formations. IET Control Theory and Applications 1, 452–460 (2007)
7. Yu, C., Hendrickx, J.M., Fidan, B., Anderson, B.D.O., Blondel, V.: Three and higher dimensional autonomous formations: Rigidity, persistence and structural persistence. Automatica 43, 387–402 (2007)
8. Hendrickx, J.M., Anderson, B.D.O., Delvenne, J., Blondel, V.: Directed graphs for the analysis of rigidity and persistence in autonomous agents systems. International Journal of Robust and Nonlinear Control 17, 960–981 (2007)
9. Olfati-Saber, R., Murray, R.M.: Graph rigidity and distributed formation stabilization of multi-vehicle systems. In: Proc. IEEE International Conference on Decision and Control, pp. 2965–2971 (2002)
10. Tanner, H.G., Pappas, G.J., Kumar, V.: Leader-to-Formation Stability. IEEE Trans. Robotics and Automation 20, 433–455 (2004)
11. Lin, Z., Francis, B., Maggiore, M.: Necessary and sufficient graphical conditions for formation control of unicycles. IEEE Trans. Automatic Control 50, 121–127 (2005)
12. Yu, C., Fidan, B., Hendrickx, J.M., Anderson, B.D.O.: Multiple formations merging: a meta-formation prospective. In: Proc. IEEE Conference on Decision and Control, pp. 4657–4663 (2006)
13. Yu, C., Fidan, B.D.O., Anderson, B.D.O.: Principles to Control Autonomous Formation Merging. In: Proc. American Control Conference, pp. 762–768 (2006)
14. Hendrickx, J.M., Yu, C., Fidan, B.: Rigidity and persistence of meta-formations. In: Proc. IEEE Conference on Decision and Control, pp. 5980–5985 (2006)
15. Hendrickx, J.M., Fidan, B., Yu, C., Anderson, B.D.O., Blondel, V.: Formation re-organization by primitive operations on directed graphs. IEEE Trans. on Automatic Control 53, 968–979 (2008)
16. Hendrickx, J.M., Yu, C., Fidan, B., Anderson, B.D.O.: Rigidity and persistence for ensuring shape maintenance of multiagent meta formations. Asian Journal of Control 10, 131–143 (2008)
17. Anderson, B.D.O., Dasgupta, S., Yu, C.: Control of Directed Formations with a Leader-First Follower Structure. In: Proc. IEEE Conference on Decision and Control New Orleans, pp. 12–14 (2007)

18. Tanner, H.G., Jadbabaie, A., Pappas, G.J.: Coordination of multiple autonomous vehicles. In: IEEE Mediterranian Conference on Control and Automation (2003)
19. Sandeep, S., Fidan, B., Yu, C.: Decentralized cohesive motion control of multi-agent formations. In: Proc. Mediterranean Conference on Control and Automation, pp. 1–6 (2006)
20. Yu, C., Fidan, B., Shames, I., Sandeep, S., Anderson, B.D.O.: Collision free coordination of autonomous multi-agent systems. In: Proc. European Control Conference, pp. 900–907 (2007)
21. Moshtagh, N., Jadbabaie, A., Daniilidis, K.: Vision-based distributed coordination and flocking of multi-agent systems. In: Proc. Robotics: Science and Systems, pp. 41–48 (2005)
22. Moshtagh, N., Jadbabaie, A.: Distributed Geodesic Control Laws for Flocking of Nonholonomic Agents. IEEE Trans. on Automatic Control 52, 681–686 (2007)
23. Moshtagh, N., Jadbabaie, A., Daniilidis, K.: Vision-based Control Laws for Distributed Flocking of Nonholonomic Agents. In: Proc. IEEE International Conference on Robotics and Automation, pp. 2769–2774 (2006)
24. Olfati-Saber, R.: Flocking for multi-agent dynamic systems: algorithms and theory. IEEE Trans. Automatic Control, 401–420 (2006)
25. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov

A Model for Cooperative Design Based on Multi-Agent System

Hua Chen, Jun Zhao, and Bo Sun

The Institute of Mechanical and Electrical Engineer,
Xi'an Technological University, 710032 Xi'an, China
zhjfylove@163.com

Abstract. For lack of information exchange and mechanism to share knowledge,the traditional distributed collaborative design mode is more difficult to adapt to the increasing development needs of product design. This paper propose an collaborative design mode based on multi-agent, according to the environment of the collaborative design and design goal, the whole process of product development is considered by this model effectively, so it will shorten the time of design, reduce the cost of product, improve quality and response the need of customers rapidly.

Keywords: CSCWD; MAS; X.509 Certificate; XML signature; XML encryption.

1 Introduction

With the universal use of the worldwide net and Internet characterized by information technology, computer supported cooperative work in design has gained more and more attentions. CSCWD make different people at different locations co-design a same product [1,2]. Because it is executed by the team who is at different geography, information exchange, knowledge sharing and exchange of the design result between the team must overcome the areal restrictions . As a result, it will bring about difficulties to access to relevant information for the user. Machine manufacture enterprise are in the process of using conventional CAD/CAM/CAE/PDM software, they often refer to queries and selection of information resources, such as a variety of standards and norms, common basic data, curve, etc, it also refer to professional application, as checking calculation and design analysis of the common mechanical parts and standard component, etc. When it refer to queries of standard basic data resources to calculate popular formula and professional check, engineer and technician still need to use traditional paper information to query, select parameters, cope result. Related parameters are input into universal CAD/CAM software system to design , calculate and draw the mechanical products , this way of working has more difficulties to adapt the need of evolving product design.Meanwhile,design data information is dispersed ,self-contained ,and cannot reflect inner link between those data material under the model of dependence on the traditional paper material.

Agent serve as a soft entity with self-governing, interactive, adaptability and intelligence ,it can help people obtain information and services independently , then finish a complicated job [3,4,5]. CAD or CAM software can be regarded as a agent.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 349–356, 2010.

They have a different communication interface with other general application program, allowing them to represent user and other agent or other person to accomplish all the tasks in collaboration. Multi agent system is composed of several agent of interactions and connection, every agent in system is capable of cooperating, cohering, consulting or mutual service and completing the tasks together [6]. Cooperative solving ability of Multi agent system is far more than a single agent. So, multi agent system is an ideal choice which can solve complex problems of multi-task, multi-object.

This paper proposes collaborative design model based on multi agent. In accordance with collaborative design environment and requirements, this model consists of five parts. These agents collaborate to achieve the following objectives:

1. Manage the product development process to avoid duplication of design, it includes managing design data, supporting the designer access to information in different design stages and releasing of information and knowledge in distributed environment.
2. Through effective communication, information access and coordination, it can help distribution of the design team Establish and maintain design process.
3. Provide content-based data exchange , information and knowledge can share and reuse in distributed design environment.
4. Provide isomeric tool software to guarantee normal work of information flow for designer.
5. Using XML data format accomplish the information alternately.
6. Using XML signature and encryption mechanisms provide security for multi-agent communication [7,8,9].

2 Cooperative Design Model for Based on Multi-Agent system

This model contains five types of agent, it is respectively design management agent (DMA), digital certificate management agent (CMA), knowledge maintenance agent (KMA), tools agent (TA), application agent (AA).

The kernel of system is design management agent (DMA), DMA mainly come to coordinating and management for the whole system activities. Agent tools include the electronic whiteboard, conference management, FTP, Email, etc; they mainly supply cooperative tools for DMA. Registering information about the agent and conflict process in the collaboration process is always transmitted and received by Internet. Agent applications include CAD / CAM system agent, control agent, CAPP system agent, etc. Developers can use their familiar tools to do this job, they collaborate with existing agent in the system by sending coordinated request to DMA.XML data formats are used to describe information and data for different applications, interacted information between agent mainly are:

1) Standard information includes the creator of information, version, creation date, modification date, etc
2) External information includes interface standards, operating parameters, configuration, etc.
3) Technical information includes the operating platform, enabling technology.
4) Performance information includes performance parameters and limitations and so on
5) Internal information includes technical details, operational logic, etc

2.1 Basic Structure of Agent

The internal structure of agent is mainly to describe the module and how these modules work together. In the collaborative design and manufacturing system based agent, Agent's features include: network center-based, interactive, semi-autonomy, responsiveness, consultative, collaborative, proactive, predictive, adaptive, flexible, durable and movable. Though, in the specific design and manufacture of agent -based system may only need a subset of these features [10], to achieve these characteristics, the appropriate module is necessary. The simple agent may only need a small number of modules (such as perception, inference, and execution), while the complex agent need more. Based on the previous model, the subject put forward the basic structure of agent, shown in Figure 1.

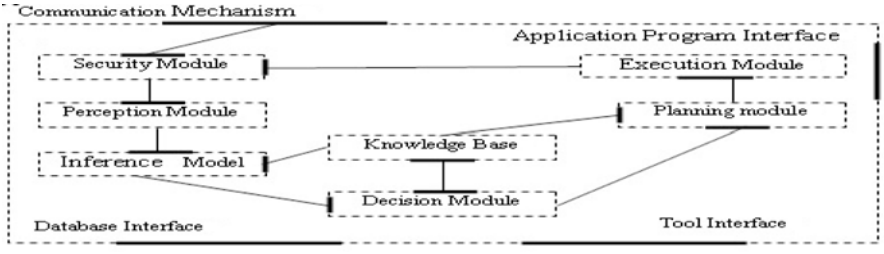


Fig. 1. General agent architecture

1) Communication rules: to send the mission, to transmit the information and express the attitude of the target tasks on various agents

2) Application programming interface, database interface and the tool interface: the interface is provided between agent and the application, database and connection of some tools.

3) Security modules: supply the security services for the interact information between the agent and the outside world, such as encryption or decryption, digital signature and signature verification [11].

4) Perception module: it is responsible for the sensation of the outside information, understanding of the situation as complete as possible. At the same time, it filter the received information, assist reasoning module identify and translate the information.

5) The reasoning module: recognize, translate, decompensate the received information .The module make agent with a higher intelligence, it is the key to the agent with the complex decision-making and knowledge processing.

6) Decision-making modules: According to the information received and the agent's goals make decisions based on the existing knowledge.

7) Planning module: according to the overall objective of agent plan their behavior.

8) The implementation of the modules: are mainly used for implementation planning.

9) Knowledge library: it contains two types of knowledge, one is the rule and the other is the knowledge block. Agents accomplish tasks conveniently and independently according to this knowledge.

Based on the basic structure of agent, the other agent’s structure will go into details.

2.2 Design Management Agent

Design management agent (Figure 2) mainly is responsible for coordination and management of the entire process of product development, it can be a people or an expert system, etc, and DMA's main functions are as follows:

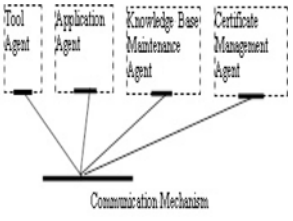


Fig.2. Design management agent

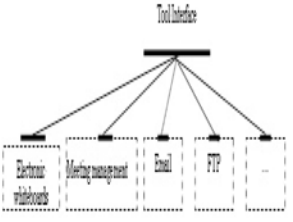


Fig. 3. Architecture of tool agent

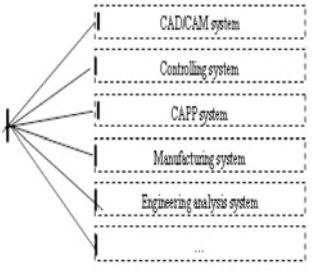


Fig. 4. Application agent

1) Manage the entire collaborative design process, including decision making, monitoring progress, assigning tasks.

2) Control collaboration and consultation between agents, understand the state of resources uses, monitor and evaluate the design process and resolute the conflict.

3) And interact with other agent systems.

DMA Knowledge library Base is composed of Task processing library, the product information database and resource library. Tasking library provides product information throughout the design process; Product information database provide all the information in the product life cycle; Resources library provides existing resources information , such as tools, applications, related agent, expert information, etc.

This knowledge can be acquired by inheritance from past experience and learning by way of. Through existing knowledge, DMA in collaboration with other agent, follow the appropriate rules, make implementation and management of this system.

2.3 Tools Agent

Most tools belong to an existing system, These systems provide very little interoperability among these single Application package , Traditional technology does not have a satisfactory solution to integrate these systems. By agent Technology , Each tool is encapsulated as an agent. Tools agent serve as middleware tool between agent and these tools including electronic whiteboards, meeting management, FTP, Email, etc, it mainly contains an interactive layer and the interface which is connected with these tools. Figure 3 shows the structure of agent tool.

2.4 Application Agent

The main role of application agent is : discover, provide and manage application based on design needs, including CAD / CAM system agent, controlling system agent, CAPP system agent, manufacturing system agent, engineering analysis system agent, etc. Its structure is similar with tools agent (Figure 4).

2.5 Database Maintenance Agent

Database maintenance agent provides a connection which is establishing between the agent and the existing project (or product) database, it is mainly for the storage and management of product data. Production data is mainly used for the corresponding data storage in the design process, engineering data is mainly used for storage national standards data in existing. The structure of the database maintenance agent is shown in Figure 5.

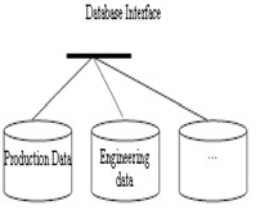


Fig. 5. Database maintenance agent

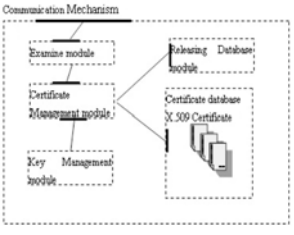


Fig. 6. Certificate management agent

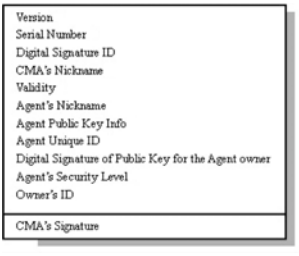


Fig. 7. Digital certificate format for agent

2.6 Digital Certificate Management Agent

Digital certificates management agent realized authentication on the agent in the form of X.509 certificate (Figure 6). In addition to general functions, Agent-MA contains the audit module, certificate management module, key management module, information dissemination module and digital certificate database.

1) Audit Module

Audit module is responsible for the checking information of the user identity and application information for review, and the request who meet the requirements is submitted to the certificate management module.

2) The certificate management module

CMA certificate management module is to provide trust services to the core service facilities. Mainly provides the following services: digital certificate system and the user is responsible for the links, that is to receive the user's service request (certificate of application, revocation and updates), and the results returned to the user. As shown in diagram 7:

3) The key management module

Key management module is responsible for management of the system key and user key. It includes key generation, updating, archiving, recovery, destruction and maintenance of key historical. The main function of the key management module is:

A, Control the key generation and storage security, supply the black box services for the certificate management.

B, The key management service, it provides the key preservation, restoration and operation of inquiry to ensure that all actions associated with user reversible.

4) Information dissemination module

News module includes the certificate database and CRL. The module's main function is to receive a certificate issued by CMA and CRL data, accept request of the certificate user query and verify. News module on the data of all certificates and CRL CMA database server related data.

5) Certificate Database

Database server for data key, user information and the processing status information in the authentication centers such as old local chronicles and statistical information storage and management. Database system uses a two-machine backup, disk arrays, etc., in order to maintain the database system stability, security, performance and scalability.

3 System Implementation

3.1 Digital Certificate

Digital certificate is a digital ID, it used to provide network identification. Simple certificate contains the public key, name, and certificate authority digital signature. Simple certificate contains the public key, name, and certificate authority digital signature. In general, certificate also contains the issuing authority (certificate authority) of the name, certificate serial number and other information..

3.1.1 Mathematical Model of X.509 Certificates

In general, X.509 v3 certificate version includes all the relevant information such as user name U, its corresponding public key $K_{pub,A}$ and the signature($Sing_{CA}(U, K_{pub}, CN, D_s, D_e)$) of the issuer. Certificate format is: $Cert_A = (ID_A \| K_{pub,A} \| S(ID_A \| K_{pub,A}, K_{priv,CA}))$

Which $K_{pub,A}$ is a effective public key of agent A between effective date D_S and the expiry date D_e

$S(ID_A \| K_{pub,A}, K_{priv,CA})$ present the signature of certification authority center. Thus, the certificate achieves binding for the agent's identity and public key $K_{pub,A}$.

3.1.2 Agent-Oriented Digital Certificate

Digital certificate is an identity proof of agent entity in the virtual world. Now including X. 509, PGP, the existing PKI certificates are all for people identification. In real life, there is basic trust relationship among people, the trust chain relationships is easy to be built up in system based on identity [12], position and role. However, in multi- agent virtual world (Figure 2), in addition to the third party certification, trust among these agent entities also depends on its owner's trust. So, when agent implement the illegal operation, to ensure that Multi agent System can track and audit liability to its owner. Agent-oriented digital certificates need to achieve banding between agent and their legitimate owners identity. Shown in Figure 7, based on X. 509 3rd edition, agent digital certificate needs to do the following modification and expansion.

1) The user name of certificate, public key information of certificate holders and unique identifier of certificate holders were changed to entity alias , public key information and unique identifier of agent. Here, the alias is the unique identifier of agent entity within the agent platform. A multi- agent system can be made up of multiple agent platforms, so in a multi- agent system, agent's unique identifier can be defined as: agent alias @ agent platform name, for example book-buyer agent @platform.com. Agent public key information, includes public key encryption algorithm identifier, key length and the agent's public key.
2) Issuing authority name, unique identifier of the issuing authority and issuing authority figures signatures were changed to agent alias of (CMA) certificate issued by the certificate management agent , global unique identifier and digital signature value. Certificate management agent is constituted by a team of agents which have very high security level security services.
3) Using signature private key of agent legal owner do signature processing for the agent's public key information, has it as investigate legal basis of its owner liability. Meanwhile, make agent's security level and the identity of the agent owner serve as extended field to add to agent digital certificate agent. It is easy to implement access controlling based on role and agent community management based on security level and running state.

3.1.3 XML Digital Signature

XML digital signatures ensure message authentication, data integrity and non-repudiation. XML digital signatures can be used in any level of the tree structure. It can be a single element a part or the entire document. The mathematical model is: $S = Sign(h(X_D), K_{SA,priv})$ .Here, S is the signature results; m represents the original message to be signed; $K_{SA,priv}$ is the sender key. Verification process is used to do verify the validity of signatures, the mathematical model is: $V = \{valid, invalid\} \leftarrow verify(X_D, S, K_{SA,pub})$ .Here, v indicate verification results, its value is valid or invalid; m indicate the information that has been signed; $K_{SA,pub}$ is the sender's public key.

3.1.4 XML Encryption

XML Encryption ensure confidentiality of the message, XML encryption allows customers to encrypt for the part of the XML document, While keep other original data unchanged. Mathematical model of encryption and decryption process is: $X_D = D(E(X_D, K_{RA,pub}), K_{RA,priv})$

Here:

X_D - indicate all the information to encrypt or decrypt text;

E - Encryption process, the input parameters is m , KSA,Pub, KSA,Pub is the receiver's public key. We can obtain the cipher text c after encryption, $c \in C$, $C \leftarrow E(X_D, K_{SA,pub})$.

D - Decryption process, the input parameters is c, KRA,Priv, $c \in C$,KRA,Priv is the recipient's private key , We can obtain the original information XD after decryption, , $X_D \leftarrow D(c, K_{priv})$.

4 Conclusions

For the shortcomings of traditional distributed collaborative design model, this paper presents an collaborative design model based multi- agent, The model can provide information services for various process activities in the product design cycle. And effectively organize multi-disciplinary product development team; Make full use of various computer-aided tools .And it can shorten product development cycles and lower costs. Meanwhile, we achieved entities certification for the communication of both through agent -oriented digital certificates in the model. The XML encryption and signature mechanisms ensure the safety of agent Interaction.

References

[1] Hua, C.: Intelligent Cooperative Design Based on Multi-agent System on Internet and System Development. In: Proceedings of International Conference on Internet Computing, Las Vegas, Nevada, USA, June 21-24, pp. 195–200 (2004)
[2] Zha, X.F.: A knowledge intensive multi-agent framework for cooperative/collaborative design modeling and decision support of assemblies. Knowledge-Based Systems 15, 493–506 (2002)
[3] Liu, H., Tang, M.: Evolutionary design in a multi-agent design environment. Applied Soft Computing 6, 207–220 (2006)
[4] Garijo, F., Tous, J., Matias Stephen Corley, J.M., Tesselaar, M.: Development of a Multi-agent System for Cooperative Work with Network Negotiation Capabilities, pp. 206–221.
[5] Shakshuki, E., Ghenniwa, H., Kamel, M.: An architecture for cooperative information systems. Knowledge-Based Systems 16, 17–27 (2003)
[6] Chira, G., Dumitrescu, D.: A Multi-agent Information Management System. CIR Report No. 042006 (June 2006) ISSN 1841 – 995X
[7] Oliveira, E., Abdelouahab, Z., Lopes, D.: Security on MASs with XML Security Specifications. In: Proceedings of the 17th International Conference on Database and Expert Systems Applications (2006)
[8] Lautenbach, B.: Introduction to XML Encryption and XML signature. Information Security Technical Report 9(3) (2004)
[9] http://www.w3.org
[10] Zhao, G., Deng, J., Shen, W.: CLOVER: an agent-based approach to systems interoperability in cooperative design systems. Computers in Industry 45, 261–276 (2001)
[11] Borselius, N.: Security in Multi-agent Systems
[12] Wenpei, H.: Multi-agent system security analysis and design. Computer Engineering and Design 28(6), 1277–1281 (2007)

Scaling Up the Accuracy of Bayesian Classifier Based on Frequent Itemsets by M-estimate

Jing Duan, Zhengkui Lin, Weiguo Yi, and Mingyu Lu

Information Science and Technology, Dalian Maritime University
Dalian, China
duanjing_1987@sina.com, dalianjx@163.com,
jiekexun98@163.com, lumingyu@tsinghua.org.cn

Abstract. Frequent Itemsets Mining Classifier (FISC) is an improved Bayesian classifier which averaging all classifiers built by frequent itemsets. Considering that in learning Bayesian network classifier, estimating probabilities from a given set of training examples is crucial, and it has been proved that m-estimate can scale up the accuracy of many Bayesian classifiers. Thus, a natural question is whether FISC with m-estimate can perform even better. Response to this problem, in this paper, we aim to scale up the accuracy of FISC by m-estimate and propose new probability estimation formulas. The experimental results show that the Laplace estimate used in the original FISC performs not very well and our m-estimate can greatly scale up the accuracy, it even outperforms other outstanding Bayesian classifiers used to compare.

Keywords: Frequent Itemsets Mining Classifier; estimating probabilities; Laplace estimate; m-estimate.

1 Introduction

A Bayesian network consists of a structure model and a set of conditional probabilities. The structural model is a directed graph in which nodes represents attributes and arcs represents attributes dependencies. But theoretically, learning an optimal Bayesian network is intractable [1]. Moreover, it has been observed that learning an unrestricted Bayesian network classifier seems to not necessarily lead to a classifier with good performance. For example, Friedman et al. [2] observed that unrestricted Bayesian network classifiers do not outperform Naïve Bayes (NB), the simplest Bayesian network classifier, on a large sample of benchmark datasets. In practice, learning restricted Bayesian network classifiers is a more realistic solution.

NB [3] is the simplest Bayesian classifier and one of the most efficient algorithms for machine learning. Despite the strong independence assumption between predictive attributes given the class value, it provides a surprisingly high level of accuracy, even compared to other more sophisticated models [4]. But in many real applications it is obvious that the conditional independence assumption in NB is rarely true, which would harm its performance in the applications with complex attribute dependencies.

During the last few years attentions have been focused on developing NB variants in order to alleviate the independence assumption between attributes, such as

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 357–364, 2010.

Decision Tree NB Hybrid (NBTree) [5], Selective Bayesian classifier (SBC) [6], Tree augmented Bayesian Network (TAN) [2], Lazy Bayesian Rules (LBR) [7], Conditional Independence Tree (CITree) [8], Boosted NB (BNB) [9], Clustered NB (CNB) [10], Extended Bayes (EB) [11], Aggregating One Dependence Estimators (AODE) [12] and so on. Among all these approaches, the Bayesian classifiers based on frequent itemsets are relatively few.

Large Bayes (LB) [13] is another outstanding Bayesian classifier which using the long itemsets. LB uses the subsets of itemset A to derive a product approximation of $P(A, c_i)$ for all classes c_i. In the training phase, LB employs an Apriori-like frequent pattern mining algorithm to discover frequent and interesting itemsets of arbitrary size together with their class supports. The class support is an estimate of the probability that the pattern occurs with a certain class label. Upon arrival of a new case to classify, a local classification model is built on the fly depending on the evidence of the new case. In the extreme case where all the discovered itemsets are of size one only, LB reduces to NB.

The group of one dependence estimators (ODEs) [14] provides a powerful alternative to NB. ODEs are very similar to NB but they also allow every attribute to depend on, at most, another attribute besides the class. Super-Parent one dependence estimators (SPODEs) [15] can be considered a subcategory of ODEs where all attributes depend on the same attribute. They have received much attention because of their efficiency in training time and their effectiveness and accuracy in classification. AODE is another outstanding Bayesian classifier. It can be considered as only using frequent itemsets whose length is 2. AODE only uses one-dependence classifiers. WAODE [16] is an improved algorithm based on AODE. Hidden One-Dependence Estimator (HODE) [17] classifier is proposed as an alternative approach to AODE in order to alleviate its problems by estimating a new variable as a super parent besides the class, whose main objective is to gather all the dependences existing in the AODE models. Hidden Naïve Bayes (HNB) [18] also creates a hidden parent for each attribute which combines the influences from all other attributes.

Frequent Itemsets Mining Classifier (FISC) [19] is an improved Bayesian classifier which averaging all classifiers built by frequent itemsets. Considering that in learning Bayesian network classifier, estimating probabilities from a given set of training examples is crucial, and it has been proved that m-estimate can scale up the accuracy of Bayesian classifiers contained NB, TAN, AODE and HNB in [20], in this paper, we aim to scale up the accuracy of FISC by m-estimate and propose new probability estimation formulas.

This paper is organized as follows. The second section presents FISC. The third section singles out a special m-estimate method after we simply introduce Laplace estimate and m-estimate. The fourth section is the experimental setting and the experimental results. The last section concludes this paper and proposes the further work.

2 FISC

Section 2.1 presents the training phase which consists of discovering the set F of all frequent itemsets. Section 2.2 shows how these itemsets and the class supports are used to classify new cases.

Assume D is a set of training instances, $A=\{A_1,\cdots,A_N\}$ is the attributes variable set, where N is the number of attributes, C is a class variable, c_i is a value of C. One training instance can be present as $\{ A_1=a_{1_j},\cdots,A_i=a_{i_j},\cdots, A_N=a_{N_j}, C=c_i \}$, where a_{i_j} is the j th value of attribute A_i.

2.1 Training

The training phase generates all the frequent itemsets using a bottom-up approach based on Apriori [21]. The input of the procedure is the database D and the output is the set of frequent itemsets F with their class counts. To facilitate the class counting, each itemset has an associated counter $count_i$ for each class c_i. Dividing a class counter by the number of tuples |D| provides the class support, that is, $l.count_i / |D| = p(l,c_i)$. An itemset is frequent, it must satisfy the requirement:

$$\frac{\sum_{i=1}^{|C|} l.count_i}{|D|} \geq minsup \ . \tag{1}$$

2.2 Classification

When given a newly observed instance $X=(A_1=a_{1_j},\cdots,A_N=a_{N_j})$ with an unknown class label, FISC firstly finds all the frequent itemsets X contains. For each frequent itemset x that X contains, FISC builds a corresponding Bayesian classifier.

Similar to other Bayesian classifier, the newly observed instance X is classified by the following equation:

$$\arg\max_{c_i} P(c_i \mid X) = \arg\max_{c_i} \frac{P(c_i,X)}{P(X)} \propto \arg\max_{c_i} P(c_i,X) \ . \tag{2}$$

Once we get a frequent itemset x ($x \subseteq X$), x divides attributes into two sets: $G_1=\{A_i \mid A_i \in x\}$ and $G_2=\{A_j \mid A_j \in X-x\}$. The attributes in G_2 are conditional independence with each other when given x and c_i, so $P(c_i,x)$ can be calculated using x by the following equation:

$$P(c_i,X)_x = P(c_i,x)P(X-x \mid c_i,x) = P(c_i,x)\prod_{a_j \in X-x} P(a_j \mid c_i,x) \ . \tag{3}$$

Finally, FISC averaged all the classifiers and classifies a new instance using (4):

$$\arg\max_{c_i \in C} \left(\frac{\sum_{x \in X} P(c_i,X)_x}{\sum_{c_i \in C}\left(\sum_{x \in X} P(c_i,X)_x\right)} \right) , \tag{4}$$

where $P(c_i,X)_x$ is computed by (3).

In order to improve the efficiency in classification, all the needed probabilities can be computed and stored in the training phase. Moreover, in order to improve the stability of the algorithm, $x = \varnothing$ can be viewed as a frequent itemset for each test sample. Therefore, in the extreme case where the test sample doesn't contain frequent itemsets, FISC reduces to NB.

3 Laplace Estimate and M-estimate

As we know, when the training examples are not enough, probabilities estimation method inevitably suffers from the zero-frequency problem. To avoid this practical problem, Laplace estimate is usually used to estimate probabilities, Just as we all know, m-estimate is another probability estimation method. In [19], FISC adopts Laplace estimate to estimate $P(c_i \mid x)$ and $P(a_j \mid c_i, x)$ as

$$P(c_i) = \frac{F(c_i)+1}{|D|+|C|}, \tag{5}$$

$$P(c_i \mid x) = \frac{F(c_i, x)+1}{F(x)+|C|}, \tag{6}$$

$$P(a_j \mid c_i, x) = \frac{F(a_j, c_i, x)+1}{F(c_i, x)+V_{A_j}}, \tag{7}$$

where $F(.)$ is the frequency with which a combination of items appears in the training data, x is a frequent itemset gained from D, V_{A_j} is the number of values of attribute A_j, $|C|$ is the number of class labels, $|D|$ is the number of training examples.

M-estimate [21] is another method to estimate probability, which can be defined as

$$P(c_i) = \frac{F(c_i)+m\times p}{|D|+m}, \tag{8}$$

$$P(a_j \mid c_i, x) = \frac{F(a_j, c_i, x)+m\times p}{F(c_i, x)+m}, \tag{9}$$

where m and p are two parameters. p is the prior estimate of the probability we wish to determine. m is a constant called the equivalent size, which determines how heavily to weight p relatively to the observed data. In fact, m-estimate can be comprehended as augmenting the actual observations by an additional m virtual samples distributed according to p. In our implementation, we set m to 1. We set p to a uniform distribution. Now, we review the equations used to estimate probabilities needed in FISC as

$$\hat{P}(c_i) = \frac{F(c_i)+1/|C|}{|D|+1}, \tag{10}$$

$$\hat{P}(c_i, x) = \frac{F(c_i, x) + 1/|C| \prod_{A_j \in x} V_{A_j}}{|D| + 1}, \tag{11}$$

$$\hat{P}(a_j | c_i, x) = \frac{F(c_i, x, a_j) + 1/V_{A_j}}{F(c_i, x) + 1}. \tag{12}$$

4 Experimental Setting and Results

We ran our experiments on 10 UCI data sets selected by WEKA [22, 23], which represent a wide range of domains and data characteristics. They are described with the number of instances, the number of class values, and the number of attributes. All of the details are described in Table 1.

Table 1. Datasets

	Domain	Size	C#	A#
1	autos	205	7	22
2	Breast-w	699	2	9
3	Car	1728	4	6
4	Diabetes	768	2	6
5	Glass	214	7	7
6	Hepatitis	155	2	16
7	labor	57	2	11
8	machine	209	7	7
9	wine	178	3	13
10	Zoo	101	7	16

In our experiments, we adopted the following three preprocessing steps.

a. Missing values: We used the unsupervised filter “Replace Missing Values” in WEKA to replace the missing values in each data set.

b. Discretization of numeric attributes: We used the supervised filter “Discretize” in WEKA to discretize all the numeric attributes.

c. Removal of useless attributes: Apparently, if the number of values of an attribute is almost equal to the number of instances in a data set, it does not contribute useful information to classification. Thus, we removed this type of attributes using the unsupervised filter “Remove useless” in WEKA.

Comparisons of the classification performances between FISC used Laplace estimate (FISC-L) and FISC used m-estimate (FISC-M) are done in terms of classification error rate. We conducted 10 runs of 10-folds CV test for comparing the classifier performances. Each datasets are divided into 10 almost equal-sized blocks randomly, and in each validation, one block was used for test data and the remaining blocks were used for training classifiers. Throughout all the tests, we measured the classification error rate, i.e., the percentage of incorrectly classified instances. To reduce the running time of FISC-L and FISC-M, we set the maximum number of the

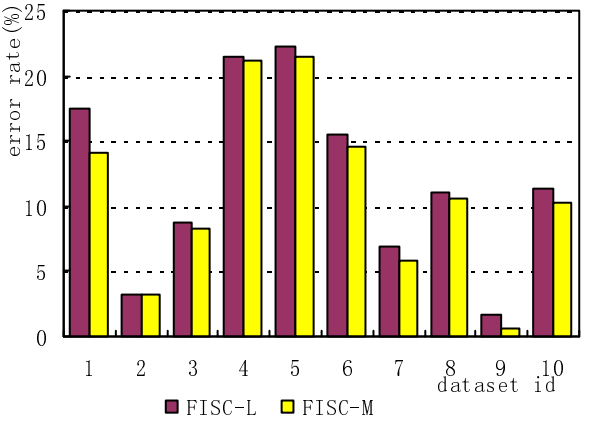


Fig. 1. Comparison between FISC-L and FISC-M

Table 2. Detiled results of error rate and standard deviation

	Domain	NB	AODE	LBR	NBTree	FISC-L	FISC-M
1	autos	26.93±2.46	18.87±1.05	20.96±1.79	20.00±1.99	17.45±0.96	14.09±1.28
2	Breast-w	2.75±0.06	3.03±0.11	2.75±0.06	2.93±0.15	3.21±0.15	3.15±0.18
3	Car	14.54±0.27	8.59±0.39	5.97±0.34	5.57±0.60	8.73±0.11	8.22±0.22
4	Diabetes	22.15±0.15	21.77±0.29	22.33±0.18	22.73±0.43	21.41±0.24	21.17±0.22
5	Glass	25.61±0.77	23.74±0.59	25.66±1.07	21.84±1.01	22.18±0.75	21.50±0.84
6	Hepatitis	15.78±0.55	14.50±0.56	14.54±1.01	15.97±1.81	15.49±0.93	14.58±0.48
7	labor	7.47±1.07	7.47±0.88	6.87±1.46	7.97±2.75	6.90±0.59	5.77±0.79
8	machine	11.49±0.38	11.29±0.72	11.73±0.33	12.77±1.41	11.10±0.65	10.64±0.45
9	wine	1.18±0.18	1.63±0.17	1.18±0.18	2.35±0.69	1.70±0.02	0.56±0.01
10	zoo	6.02±1.06	5.34±0.70	6.02±1.06	6.44±1.14	6.07±1.30	3.76±0.75
Mean		13.39±0.69	11.62±0.55	11.80±0.75	11.86±1.20	11.42±0.57	10.34±0.52

generated frequent itemsets is 1000 for all data sets. Fig. 1 shows the comparison of FISC-L and FISC-M, and the detailed experimental results are shown in Table 2.

The above experiments show that the classifiers with our m-estimate perform overall better than classifiers with Laplace estimate, i.e. FISC-M significantly outperforms FISC-L. The bars of FISC-L are almost higher than those of FISC-M.

Further, we also compare FISC-M with some other outstanding classifiers such as NB, LBR, AODE and NBTree. We implemented the experiments within the WEKA framework and used the implementation of those four classifiers in WEKA. We also conducted ten runs of 10-folds CV test for comparing the classifier performances. Table 2 shows the detailed experimental results.

We also compared related algorithms via two tailed t-test with a 95 percent confidence level. According to the statistical theory, we speak of two results for a data set as being “significantly different” only if the probability of significant difference is at least 95 percent. Table 3 shows the compared result of two-tailed t-test, and we

perform a win/draw/loss summary to compare overall perform of FISC-M against each other algorithms. For each pair-wise comparison we present first the number of data sets for which FISC-M obtained significantly lower average error than the comparator algorithm, the number for which the algorithms obtained the same average error, and the number for which the alternative algorithm obtained the significantly lower average error. From our experiments we can see that our FISC-M can rival those outstanding classifiers in performance.

Table 3. The compared results of two-tailed t-test

	Domain	NB	AODE	LBR	NBTree	FISC-L
1	autos	win	win	win	win	win
2	Breast-w	loss	draw	loss	loss	draw
3	Car	win	loss	loss	loss	win
4	Diabetes	win	win	win	win	win
5	Glass	win	win	win	draw	win
6	Hepatitis	win	draw	draw	win	win
7	labor	win	win	win	draw	win
8	machine	win	win	win	win	win
9	wine	win	win	win	win	win
10	Zoo	win	win	win	win	win
W/D/L		9/0/1	7/2/1	7/1/2	6/2/2	9/1/0

5 Conclusions

Considering that in learning Bayesian network classifier, estimating probabilities from a given set of training examples is crucial, and it has been proved that m-estimate can scale up the accuracy of Bayesian classifiers contained NB, TAN, AODE and HNB. Based on the above achievements, in this paper, we proposed new probability estimation formulas called FISC-M and aimed to scale up the accuracy of FISC. Our experimental results show that the classifiers with our m-estimate perform better than the ones with Laplace estimate used in [19]. Moreover, the proposed FISC-M outperforms other outstanding algorithms used to compare in terms of performance accuracy.

To speak frankly, FISC-M will lead to heavy time overhead and storage burden since it ensembles a large number classifiers generated by frequent itemsets. At this point, it has the same shortage with the original FISC. So how to select the optimal classifiers to ensemble is one direction for our future research. Moreover, we want to further combine decision tree with FISC-M to determine optimal amount of frequent itemsets, i.e., use different number of frequent itemsets for different datasets.

References

1. Chickering, D.M.: Learning Bayesian Networks is NP-Complete. In: Fisher, D., Lenz, H. (eds.) Learning from Data: Artificial Intelligence and Statistics, pp. 121–130. Springer, Heidelberg (1996)
2. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Mach. Learn. 29, 131–163 (1997)
3. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley & Sons Inc., Chichester (1973)

4. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In: 13th International Conference on Machine Learning, pp. 105–112. Morgan Kaufmann, Italy (1996)
5. Kohavi, R.: Scaling Up the Accuracy of Naïve Bayes Classifiers: A Decision-Tree Hybrid. In: Proc. KDD 1996, pp. 202–207 (1996)
6. Ratanamahatana, C.A., Gunopulos, D.: Scaling up the Naïve Bayesian Classifier: Using Decision Trees for Feature Selection. In: Proc. Workshop Data Cleaning and Preprocessing (DCAP 2002), at IEEE Int'l Conf. Data Mining, ICDM 2002 (2002)
7. Zheng, Z., Webb, G.I.: Lazy Learning of Bayesian Rules. Mach. Learn. 41(1), 53–84 (2000)
8. Zhang, H., Su, J.: Conditional independence trees. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 513–524. Springer, Heidelberg (2004)
9. Diao, L., Hu, K., Lu, Y., Shi, C.: A method to boost Naive Bayesian classifiers. In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002. LNCS (LNAI), vol. 2336, pp. 115–122. Springer, Heidelberg (2002)
10. Vilalta, R., Rish, I.: A decomposition of classes via clustering to explain and improve Naïve Bayes. In: Proceedings of the 14th European conference on machine learning (2003)
11. Rosell, B., Hellerstein, L.: Naïve Bayes with higher order attributes. In: Tawfik, A.Y., Goodwin, S.D. (eds.) Canadian AI 2004. LNCS (LNAI), vol. 3060, pp. 105–119. Springer, Heidelberg (2004)
12. Webb, G.I., Boughton, J., Wang, Z.: Not So Naïve Bayes: Aggregating One-Dependence Estimators. Mach. Learn. 58, 5–24 (2005)
13. Meretakis, D., Wuthrich, B.: Extending Naïve Bayes classifiers using long itemsets. In: Proceedings of the 5th ACM SIG-KDD international conference on knowledge discovery and data mining, pp. 165–174 (1999)
14. Sahami, M.: Learning limited dependence Bayesian classifiers. In: 2nd International Conference on Knowledge Discovery in Databases, Portland, pp. 335–338 (1996)
15. Yang, Y., Korb, K., Ting, K.M., Webb, G.I.: Ensemble Selection for SuperParent-One-Dependence Estimators. In: Zhang, S., Jarvis, R.A. (eds.) AI 2005. LNCS (LNAI), vol. 3809, pp. 102–112. Springer, Heidelberg (2005)
16. Jiang, L., Zhang, H.: Weightily averaged one-dependence estimators. In: Yang, Q., Webb, G. (eds.) PRICAI 2006. LNCS (LNAI), vol. 4099, pp. 970–974. Springer, Heidelberg (2006)
17. Flores, M.J., Gamez, J.A., Martinez, A.M., Puerta, J.M.: HODE: Hidden One-Dependence Estimator. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS, vol. 5590, pp. 481–492. Springer, Heidelberg (2009)
18. Jiang, L., Zhang, H., Cai, Z.: A Novel Bayes Model: Hidden Naïve Bayes. IEEE Transactions on Knowledge and Data Engineering 21(10) (2009)
19. Xu, J., Jiang, Y., Zhou, Z.: Bayesian Classifier Based on Frequent Item Set Mining. Journal of Computer Research and Development 44(8), 1293–1300 (2007)
20. Jiang, L., Wang, D., Cai, Z.: Scaling up the accuracy of Bayesian Network classifier by M-estimate. In: Huang, D.-S., Heutte, L., Loog, M. (eds.) ICIC 2007. LNCS (LNAI), vol. 4682, pp. 475–484. Springer, Heidelberg (2007)
21. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
22. Witten, I.H., Frank, E.: Data mining: Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann, San Francisco (2000), `http://prdownloads.sourceforge.net/weka/datasets-UCI.jar`
23. Merz, C., Murphy, P., Aha, D.: UCI repository of machine learning databases. Dept of ICS, University of California, and Irvine, CA

Aggressive Dimensionality Reduction with Reinforcement Local Feature Selection for Text Categorization

Wenbin Zheng and Yuntao Qian

College of Computer Science, Zhejiang University
310027, Hangzhou, China
zwb@zju.edu.cn

Abstract. A major problem of text categorization is the high dimensionality of the input feature space. This paper proposes a novel approach for aggressive dimensionality reduction in text categorization. This method utilizes the local feature selection to obtain more positive terms and then scales the weighting in the global level to suit the classifier. After that the weighting is enhanced with the feature selection measure to improve the distinguishing capability. The validity of this method is tested on two benchmark corpuses by the SVM classifier with four standard feature selection measures.

Keywords: Text Categorization; Dimensionality Reduction; Local; Feature Selection; Reinforcement.

1 Introduction

Text categorization (TC) is a task of automatically assigning predefined categories to a given text document based on the content[1]. A growing number of machine learning techniques have been applied to the TC in recent years. For example, k-nearest neighbor (kNN)[2], neural networks[3], support vector machines (SVM)[4][5] and concept-based model[6].

In the TC, the high dimensionality of the input feature space is still a major difficulty. Generally, there are tens of thousands or even hundreds of thousands of dimensions. However, such high dimensionality is not permitted for many learning algorithms. Moreover most of these dimensions are not relative to the text categorization and their impact on increasing complexity is worse relatively[2]. So it is necessary to reduce the dimensionality of the feature space. Features selection (FS) is one of the most used methods for the purpose of dimensionality reduction by selecting significant terms, which minimizes the loss of performance and maximizes the reduction of complexity.

It can be performed basically in two ways: local and global FS[1]. For the local FS, a set of terms is selected based on the relevant and irrelevant documents in the category, each category is represented by a set of unique features. In the global FS, a set of terms is chosen for all the categories based on the

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 365–372, 2010.

value of each term-assessed category-independent. In most reports, the number of dimensionality can be reduced more than 90%[2]. More aggressive FS, with more than 95% reduction, is reported in[7] [8].

It should be noted that there are some problems in aggressive dimensionality reduction with the FS. For the global FS, averaging the score of a feature will reduce significantly the high relevancy of the term to the specific category as the score might be dispersed to other category[9]. So the number of features may become zero after aggressive FS, and the accuracy of classifier will be dropped drastically. For the local FS, although it manages to preserve better specific category information, the number of features being selected is also small after aggressive FS. So the capability of distinguishing categorization is impacted easily by one or two dominance terms.

To overcome these problems above, this paper proposes a reinforcement local FS approach, in which more than 99% dimensionality reduction is implemented and at the same time the accuracy of TC has few loss even a few improvement.

This method utilizes the character that local FS dose not take into account the negative features to obtain more positive terms after aggressive dimensionality reduction. Then the weighting of these selected terms is scaled in a global level in order to suit the classifier in which the discrimination function value can be compared in the same scale. After that these weighting values are weighted with the FS measure to enhance the capability of distinguishing categorization.

The rest of this paper is organized as follows: Section 2 explains the proposed approach in detail. Experimental results and analyses are given in Sections 3. Conclusion is presented in Sections 4.

2 The Reinforcement Local Feature Selection

2.1 The Feature Selection Measure

The FS measure is used as the criterion of term space reduction. Here, we present four commonly known FS measures (see Table 1). To be brief, their mathematical justifications are omitted and more descriptions can be found in[1].

In Table 1, each of these measures is actually a function $f(c_k, t_k)$ with a term t_k and a category c_i as its parameters. The function value indicates some

Table 1. Feature selection measures

Description	Equation
Information Gain	$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log \frac{P(t,c)}{P(t)P(c)}$
CHI-Square	$CHI(t_k, c_i) = \frac{N[P(t_k,c_i) \cdot P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i) \cdot P(\bar{t}_k,c_i)]^2}{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}$
Correlation Coefficient	$CC(t_k, c_i) = \frac{\sqrt{N}[P(t_k,c_i) \cdot P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i) \cdot P(\bar{t}_k,c_i)]}{\sqrt{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}}$
GSS-Coefficient	$GSS(t_k, c_i) = P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)$

relationship between the term and the category. In the global FS, the value of a term is assessed in a global, or category-independent. Either the average or the maximum of their category-specific values is computed[2]. That is,

$$f_{avg}(t) = \sum_{i=1}^{m} P(c_i) f(t, c_i) \tag{1}$$

$$f_{\max}(t) = \max_{i=1}^{m} \{f(t, c_i)\} \tag{2}$$

2.2 The Local Feature Selection and Weighing

In the local FS, a term is assessed with a FS measures for each category, and all the terms are sorted according to the measure value. Then for each category, a set of terms with high measure value is selected as the features and the rest of terms are removed. Therefore when a document belonging to a specific category is used, the chance of significant terms appearing to classifier for the specific category is increased because the local FS does take into account of the positive features.

Here, given a document d_j, its feature (term) vector is represented by

$$d_j = (t_1, t_2, \cdots, t_N) \tag{3}$$

where N is the number of native features.

We calculate the weighting of term t_k in document d_j with the function $tfidf(t_k, d_j) = \#(t_k, d_j) \cdot idf(t_k)$, where $\#(t_k, d_j)$ denotes the number of times t_k occurred in d_j, and $idf(t_k)$ is the inverse documents frequency which is defined as

$$idf(t_k) = \log \frac{|Tr|}{\#(t_k)} \tag{4}$$

where $\#(t_k)$ denotes the number of documents in training set in which t_k occurs at least once, and $|Tr|$ denotes the number of documents in training set.

For a specific category c_i from the category set space $\{c_1, c_2, \cdots, c_m\}$, we select a set of terms by scoring the value of the FS measure function $f(t_k, c_i)$ (see Table 1), and then the document d_j can be represented by

$$d_j \Rightarrow \begin{cases} d_j^1 = (\tau_1^1, \tau_2^1, \cdots, \tau_n^1) \\ \vdots \\ d_j^i = (\tau_1^i, \tau_2^i, \cdots, \tau_n^i) \quad n << N \\ \vdots \\ d_j^m = (\tau_1^m, \tau_2^m, \cdots, \tau_n^m) \end{cases} \tag{5}$$

where d_j^i means the document d_j after FS for a specific category c_i, and $(\tau_1^i, \tau_2^i, \cdots, \tau_n^i)$ is a subset of $(t_1, t_2, \cdots, t_N)$. So every term τ_i^j has a weighting value $tfidf(\tau_k^i, d_j)$.

In order to let the $tfidf(\tau_k^i, d_j)$ falls in the [0,1) interval, the weighting value needs to be scaled. However, normalization only in the local level would lead to different scale because different category space has different feature set. Therefore, we scale the feature weighting value in the global level so that the discrimination function value calculated by classifier can be compared in the same scale. Then the weighting of term τ_k^i in document d_j is calculated as

$$\omega(\tau_k^i, d_j) = \frac{tfidf(\tau_k^i, d_j)}{\sqrt{\sum_{s=1}^{|N|} (tfidf(t_s, d_j))^2}} \tag{6}$$

where i is from 1 to n, $n << N$.

To avoid the disadvantage impact of dominance terms, we multiply the weighting value with the FS measure function. So the effect of a dominance term with low relevancy will descend and the effect of a dominance term with high relevancy will increase. Then the capability of distinguishing categorization will be improved. We obtain the following equation

$$w(\tau_k^i, d_j) = \frac{tfidf(\tau_k^i, d_j) \cdot f(\tau_k^i, c_i)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))^2}} \tag{7}$$

Finally, a document d_j can be represented by

$$d_j \Rightarrow \begin{cases} d_j^1 = (w_1^1, w_2^1, \cdots, w_n^1) \\ \vdots \\ d_j^i = (w_1^i, w_2^i, \cdots, w_n^i) \\ \vdots \\ d_j^m = (w_1^m, w_2^m, \cdots, w_n^m) \end{cases} \tag{8}$$

where $w_k^i = w(\tau_k^i, d_j)$.

2.3 The Categorization Algorithm

We apply the linear SVM with the one-versus-rest method as the classifier. The more descriptions about the SVM theory can be referred to [4]. A pseudo-code of our TC algorithm is given in Algorithm 1.

3 Experiment

3.1 Experimental Setting

We use two standard TC benchmarks in this evaluation: Reuters-21578 and 20 Newsgroups. It can be downloaded from a publicly available repository of datasets[1]. For the Reuters-21578 dataset, we use the files r8-train-stemmed and

[1] Available at http://web.ist.utl.pt/~acardoso/datasets/

Algorithm 1. The TC algorithm with the reinforcement local FS

Input: A set of labeled vectors of the training and test set
Output: A set of labels of the test set given by classifier
for all t_k appearing in the native term space $(t_1, t_2, \cdots, t_N)$ **do**
calculating $idf(t_k)$ using the equation 4
for all c_i in the category space $\{c_1, c_2, \cdots, c_m\}$ **do**
calculating the FS measure $f(t_k, c_i)$ in Table 1
end for
end for
for all d_j in training set **do**
for all c_i in the category space $\{c_1, c_2, \cdots, c_m\}$ **do**
calculating the d_j using the equation 5 6 7 8
using d_j as input of classifier for SVM leaning
end for
end for
for all d_j in test set **do**
for all c_i in the category space $\{c_1, c_2, \cdots, c_m\}$ **do**
calculating the d_j using the equation 5 6 7 8
calculating the output value of SVM classifier using d_j as input
end for
assigning a category label to d_j according to the output value of the classifier
end for

r8-test-stemmed, and for the 20 Newsgroups, we use the files 20ng-train-stemmed and 20ng-test-stemmed. These files have been pre-processed including the removal of short words, removal of stop words, and stemming.

To evaluate the reinforcement FS method (abbreviated as RFS for convenience), we retest the norm FS method [2] (abbreviated as NFS) for comparisons.

According to the experimental results of the NFS method, the performance used in equation 2 is better than that used in equation 1. So we choose the result used in equation 2 as the result of the NFS to compare with our RFS method to be brief.

We apply the SVMlight6[2] as the SVM classifier and adopt its default parameters for comparison convenience.

3.2 Performance Measure

For a specific category c_i from the category space $\{c_1, c_2, \cdots, c_m\}$, a well-known $F1$ measure is defined as follow

$$F1_i = \frac{2 \times Pr_i \times Re_i}{Pr_i + Re_i} \tag{9}$$

where Pr_i is the precision and Re_i is the recall [1].

[2] Available at http://svmlight.joachims.org/

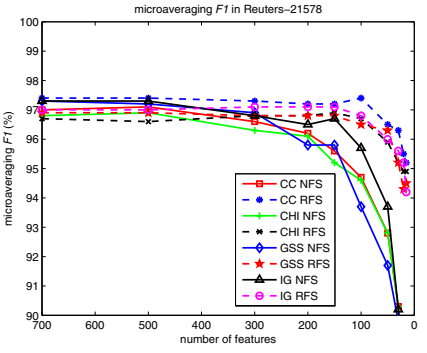


Fig. 1. The *microaveragingF*1 for four FS measures in Reuters-21578 (The legend "CC NFS" denotes the result of NFS method with the CC FS measure, and the rest may be deduced by analogy)

For all categories, we adopt the *microaveragingF*1 and the *macroaveragingF*1 to evaluate the performance overall, they are defined as follows

$$microaveragingF1 = \frac{2 \times \hat{Pr}^{\mu} \times \hat{Re}^{\mu}}{\hat{Pr}^{\mu} + \hat{Re}^{\mu}} \tag{10}$$

$$macroaveragingF1 = \frac{\sum_{i=1}^{m} F1_i}{m} \tag{11}$$

where $\hat{Pr}^{\mu}$ is the microaveraging precision and $\hat{Re}^{\mu}$ is the microaveraging recall, and their detail descriptions can be found in [1].

3.3 The Experimental Results and Analyses

The experimental results in Reuters-21578 with the FS level form 5% to 1‰ are plotted in Fig.1 and Fig.2. There is little difference about the performance between the NFS (plotted by the solid line) and RFS (plotted by the dashed line) in the 5% FS level (the number of features equal to 700). However, the performance of the NFS descends drastically while the FS level is less than 1%. On the other hand, the performance variety of the RFS is not obvious, and the best performance according to some FS measures appears mostly in about 1% FS level.

The experimental results in 20 Newsgroups with the FS level form about 5% to 1‰ are plotted in Fig.3 and Fig.4. From the two figures, we can see that the performance of NFS (plotted by the solid line) begins to drop drastically from the 3% FS level (the number of features equal to 1500). On the contrary, the performance of RFS (plotted by the dashed line) alters rarely in the same FS level.

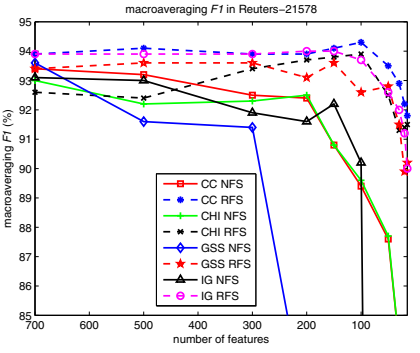


Fig. 2. The *macroaveragingF*1 for four FS measures in Reuters-21578

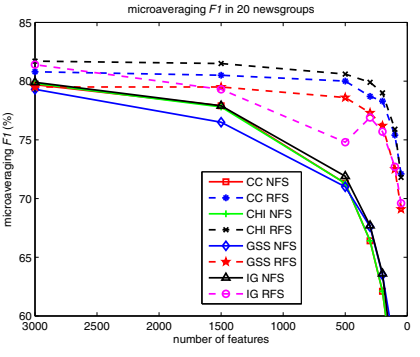


Fig. 3. The *microaveragingF*1 for four FS measures in 20 Newsgroups

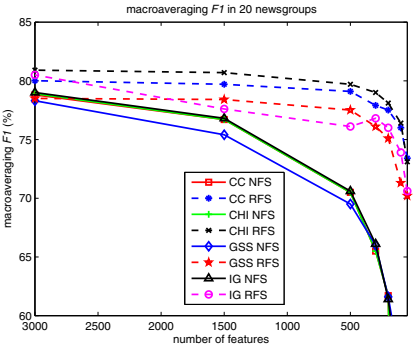


Fig. 4. The *macroaveragingF*1 for four FS measures in 20 Newsgroups

From above results, we can see that the RFS method has outperformed the NFS method in most FS measures and most FS levels especially in the situation of aggressively FS.

4 Conclusion

This paper proposes a reinforcement local FS approach for aggressive dimensionality reduction in the TC. It can accelerate the speed of the classifier as well as improve the performance after aggressive dimensionality reduction. The experimental results with four standard FS measures show the effectiveness of the proposed approach on two benchmark corpuses. The major contributions of this paper include (1) incorporating the local FS and global scaling to suit the classifier and (2) using the FS measure to enhance the distinguishing capability after the aggressive dimensionality reduction. For the further research, we are trying to incorporate sematic conception to improve the efficiency and expand the reinforcement local FS to more classifiers.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 60872071).

References

1. Sebastiani, F.: Machine learning in automated text categorization. Acm Computing Surveys 34(1), 1–47 (2002)
2. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Machine Learning-International Workshop Then Conference, pp. 412–420 (1997) (Citeseer)
3. Hwee Tou, N., Wei Boon, G., Kok Leong, L.: Feature selection, perceptron learning, and a usability case study for text categorization. In: SIGIR Forum, spec. issue., pp. 67–73 (1997)
4. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
5. Wang, T., Chiang, H.: One-against-one fuzzy support vector machine classifier: An approach to text categorization. Expert Systems with Applications 36, 10 030–10 034 (2009)
6. Shehata, S., Karray, F., Kamel, M.: A concept-based model for enhancing text categorization. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 629–637. ACM, New York (2007)
7. Makrehchi, M., Kamel, M.S.: Text classification using small number of features. In: 4th International Conference on Machine Learning and Data Minining in Pattern Recognition, pp. 580–589 (2005)
8. Gabrilovich, E., Markovitch, S.: Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4. 5. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 41. ACM, New York (2004)
9. How, B.C., Kiong, W.T.: An examination of feature selection frameworks in text categorization. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.-H. (eds.) AIRS 2005. LNCS, vol. 3689, pp. 558–564. Springer, Heidelberg (2005)

3D Shape Representation Using Gaussian Curvature Co-occurrence Matrix

Kehua Guo

School of Information Science & Engineering,
Central South University, Changsha, China
guokehua@yahoo.com.cn

Abstract. Co-occurrence matrix is traditionally used for the representation of texture information. In this paper, the co-occurrence matrix is combined with Gaussian curvature for 3D shape representation and a novel 3D shape description approach named Gaussian curvature co-occurrence matrix is proposed. Normalization process to Gaussian curvature co-occurrence matrix and the invariants independence of the translation, scaling and rotation transforms are demonstrated. Experiments indicate a better classification rate and running complexity to objects with slight different shape characteristic compared with traditional methods.

Keywords: Gaussian curvature; Co-occurrence matrix; Differential Geometry; Pattern Recognition.

1 Introduction

3D shape matching is a fundamental issue in 3D computer vision with many applications, such as shape registration, 3D object recognition and classification [1,2,3]. In recent decades, there has been lots of research on 3D shape matching. The key question in shape matching is the shape characteristic representation. Different approaches include Fourier descriptor representations [4], moment invariants representations [5], curvature-based representations [6], shape distributions [7], harmonic shape images [8], and neural network approaches [9], etc. However, many shape characteristic representations using local shape signatures cannot represent the relationship between different parts on the object, and methods based on global shape characteristic cannot distinguish objects with slight different shape characteristic.

In this paper, we propose a new solution that does not suffer from such problems. According to differential geometry principle, the inherent characteristic of a spatial surface can be represented by spatial curvature of each point [10], such as mean curvature and Gaussian curvature. A good distinguish result will be achieved even to the objects with slight different shape characteristic. But spatial curvature can only describe the local characteristic of the surface, and cannot perform well in the presence of noise, occlusion and clutter.

Gray level co-occurrence matrix, one of the most known texture analysis methods, can estimate image properties related to second-order statistics [11]. Since the gray level co-occurrence matrix can describe the gray statistical characteristic on a texture

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 373–380, 2010.

image, a new defined "Gaussian curvature co-occurrence matrix" can also represent Gaussian curvature statistical characteristic on a 3D object. Consequently the matching problem of 3D shapes can be simplified to the comparison of the new defined Gaussian curvature co-occurrence matrix.

We extend the ideas of traditional co-occurrence matrix put forward in [11] and introduce a different solution to match the 3D object with slight different shape characteristic. Traditional co-occurrence matrix has been proven to be very powerful for texture analysis, and dealt mainly with gray level to describe the characteristic of texture; here we compute the co-occurrence matrix based on the Gaussian curvature of each pixel on the spatial curved surface. Then a new shape description approach named Gaussian curvature co-occurrence matrix is proposed. Normalization process to Gaussian curvature co-occurrence matrix and some characteristic invariants independent of the translation, scaling and rotation transforms are illustrated. Finally the surfaces can be matched based on these characteristic invariants. In fact, co-occurrence matrix has been combined with the mean curvature for shape representation[12], so the idea proposed in this paper is an improvement to the mean curvature co-occurrence matrix. Experiments indicate a better classification rate and running complexity than traditional approaches to 3D objects with slight different shape characteristic.

2 Computing the Gaussian Curvature

According to differential geometry principle, curvature is the inherent characteristic of a spatial surface. When a rigid transform is applied to a spatial surface, curvature is an invariant [10]. Therefore, spatial curvature can be employed for the representation of a spatial surface. In 3D Euclidean space, given a parametric surface defined as:

$$S(x,y)=[x\ y\ f(x,y)]^T,(x,y)\in D \tag{1}$$

where $X-Y$ is the reference plane in 3D space, D is projection region of the surface to $X-Y$ plane, $f(x,y)$ represents the distance from the surface to point (x,y) in reference plane.

$S(x,y)$ can be represented by two fundamental forms. The first fundamental form can be posed as follows [10]:

$$I(dx,dy)=dS\bullet dS=(S_x dx+S_y dy)\bullet(S_x dx+S_y dy)=Edx^2+2Fdxdy+Gdy^2 \tag{2}$$

where E, F, G are parameters of the first fundamental form:

$$E=S_x\bullet S_x, F=S_x\bullet S_y, G=S_y\bullet S_y \tag{3}$$

The second fundamental form can be represented as follows [10]:

$$II(dx,dy)=-dS\bullet dn=(S_{xx}dx^2+2S_{xy}dxdy+S_{yy}dy^2)\bullet n=Ldx^2+2Mdxdy+Ndy^2 \tag{4}$$

where L, M, N represent the parameters of the second fundamental form, n is unit normal vector in point $(x,y,S(x,y))$:

$$L = S_{xx} \bullet n, M = S_{xy} \bullet n, N = S_{yy} \bullet nn = (S_x \times S_y) / | S_x \times S_y | \tag{5}$$

The two fundamental forms of a surface can be uniquely determined by six parameters: E , F , G , L , M , N .

Gaussian curvature K can be formulated as follows [13]:

$$K = \frac{LN - M^2}{EG - F^2} \tag{6}$$

For a discretized parametric surface, S_x , S_y , S_{xx} , S_{xy} , S_{yy} can be computed as follows:

$$S_x = [1\,0\,f_x]^T, S_y = [0\,1\,f_y]^T, S_{xx} = [0\,0\,f_{xx}]^T, S_{xy} = [0\,0\,f_{xy}]^T, S_{yy} = [0\,0\,f_{yy}]^T \tag{7}$$

So that Gaussian curvature K can be computed according to following formulas:

$$K = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \tag{8}$$

For digital range image surface, approximations can be computed by local polynomial fitting approach, $n \times n$ operator is usually utilized to the convolution operation with original range image [14]:

$$f_x = D_x * f, f_y = D_y * f, f_{xx} = D_{xx} * f, f_{xy} = D_{xy} * f, f_{yy} = D_{yy} * f \tag{9}$$

where D is $n \times n$ operator. Paul et al. [14] have proposed the formulas for $n = 7$:

$$D_x = D_0 D_1^T, D_y = D_1 D_0^T, D_{xx} = D_0 D_2^T, D_{yy} = D_2 D_0^T, D_{xy} = D_1 D_1^T \tag{10}$$

where

$$D_0 = \frac{1}{7}[1\ 1\ 1\ 1\ 1\ 1\ 1]^T, D_1 = \frac{1}{28}[-3\ -2\ -1\ 0\ 1\ 2\ 3]^T, D_2 = \frac{1}{84}[5\ 0\ -3\ -4\ -3\ 0\ 5]^T \tag{11}$$

where D_0 , D_1 , D_2 are column vectors for window operator computing.

3 Gaussian Curvature Co-occurrence Matrix

3.1 Construction of Gaussian Curvature Co-occurrence Matrix

The surface of 3D objects can be considered to be composed of spatial pixel points characterized by spatial curvature. Spatial distribution of the Gaussian curvatures discriminates different shape classes. Traditional co-occurrence matrix can well describe the texture characteristic, and it is computed based on gray level. In this paper, the co-occurrence matrix will be compute based on “Gaussian curvature level”,

and then we will construct some invariants, which are independent of the translation, scaling and rotation transforms.

To measure the relationship of gray level on the texture image, co-occurrence matrix has been introduced [11]. Let $f(x, y)$ be an image of size $M \times N$. Suppose that the gray value of each pixel is normalized into n levels. A $d-$dependent co-occurrence matrix may be defined as a two-dimensional array whose generic element $p_d(i, j)$ represents the joint probability, approximated by the relative frequency, of the occurrence of a pair of points, spatially separated by d pixels, one having gray level i and the other with gray level j.

In our work, co-occurrence matrix will be constructed based on Gaussian curvature value of each pixel. Firstly, we will normalize all the Gaussian curvature values into n levels, then define a $d-$dependent co-occurrence matrix whose element $p_d(i, j)$ represents the occurrence of a pair of points, spatially separated by d pixels, one having Gaussian curvature level i and the other with level j.

However, in many cases, a lot of plane points, whose Gaussian curvature are zero, can exist in the surface. They cannot contribute to the matching but will occupy large computing and reduce the matching efficiency. So that these plane points will be discarded before matching. Therefore, the definition of Gaussian curvature co-occurrence matrix is proposed as follows:

Definition 1: Let S be a spatial curved surface, a Gaussian Curvature Co-occurrence Matrix named $p_d(i, j)$ is defined as:

$$P_d(i,j) = \#\{p_1, p_2 \in S \mid K(p_1) = i \neq 0, K(p_2) = j \neq 0, D(p_1, p_2) = d\} \quad (12)$$

where $D(p_1, p_2)$ is the Euclidean distance between point p_1 and p_2, $\#$ represents the cardinality of a set.

Co-occurrence matrix is quite effective for discriminating different textures; however, it has the disadvantage of a high computational cost. F. Argenti et al. [15] presented fast algorithms for texture analysis using co-occurrence matrix, these algorithms will be applied in our approach to shape analysis.

3.2 Normalization and Invariants

In Definition 1, $p_d(i, j)$ represents the occurrence of a pair of points, spatially separated by d, one having Gaussian curvature level i and the other with level j. After a rotation transform, the spatial relationship and Gaussian curvature level of each pair will not be changed. Therefore, Gaussian curvature co-occurrence matrix will be independent to rotation transform.

We consider a scaling transform with a scaling factor. After this transform, Gaussian curvature of every pixel point will be changed according to the scaling factor. So the normalized Gaussian curvature level value n is invariant. In addition,

the size of neighborhood d must change with the scaling factor. In our approach, we define the value of d as follows:

$$d = k\sqrt{|S|} \tag{13}$$

where $k > 0$ is a neighborhood factor

Hence, the independence of the Gaussian curvature co-occurrence matrix to scaling transforms can be guaranteed by neighborhood factor k .

Measure of co-occurrence matrix uniformity may be used for discrimination. Several parameters have been proposed to analyze texture in [11]. To Gaussian curvature co-occurrence matrix, we define four invariants to describe the shape characteristic of 3D objects:

$$\begin{cases} CON: T_1 = \sum (i-j)^2 P_d(i,j) \\ ASM: T_2 = \sum P_d^2(i,j) \\ ENT: T_3 = -\sum P_d(i,j) \lg P_d(i,j) \\ COR: T_4 = \dfrac{1}{\sigma_x \sigma_y} (\sum ijP_d(i,j) - \mu_x \mu_y) \end{cases} \tag{14}$$

where μ_x and σ_x are the Gaussian and standard deviation of the row sums of the Gaussian curvature co-occurrence matrix, and μ_y and σ_y are analogous statistics of the column sums.

4 Experimentation Results

The classification performance of Gaussian curvature co-occurrence matrix will be verified in the first experiment. 15 images with slight different shape characteristic are shown in Fig. 1. The sample objects in the gallery are numbered from 1 to 15. This gallery consists of 10 different kinds of shapes: {1}, {2,11}, {3,10}, {4,12}, {5,8}, {6}, {7}, {9,13}, {14} and {15}. The scene size of each object is 240 pixels. In this experiment,some traditional approaches for classification and our algorithm will be applied in classification process for performance comparison.

To each object, we firstly compute the Gaussian curvature of each pixel on the 3D surface. Secondly, a neighborhood factor k will be selected to construct the Gaussian curvature co-occurrence matrix according to Definition 1. Again, the shape characteristic invariants are computed. Finally, classification process will be performed through measuring the similarity distance to every pair of objects in the gallery. In our experiment, one-order Minkowski distance is employed to measure the difference of each two surfaces in the gallery. Let S and S' be two spatial curved surface, the one-order Minkowski distance between S and S' is defined as follows:

$$DIF(S,S') = \frac{1}{4}\sum_{i=1}^{4} \frac{|T_i - T'_i|}{\min(T_i, T'_i)} \tag{15}$$

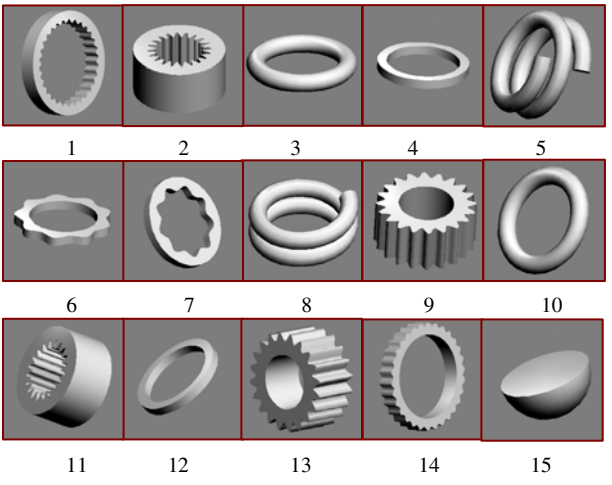


Fig. 1. 15 images with slight difference

In our algorithm, a good neighborhood factor is chosen as 0.4 for the performance comparison(the choosing approach of neighborhood factor is proposed in reference[12]), and the threshold value is 20%. In addition, we apply some traditional approach proposed in [4,5,6,12] for comparision. The classification results of applying these approaches are shown in the Tab.1. Classification rate represents the percentage of correct classification quantity in the gallery. The fourth column of Tab.1 illustrates the running time of every algorithm.

Table 1. Classification result(CPU: PIV 2.0G Hz, RAM: 512M)

Algorithm	Classification Result	Classification Rate	Time (ms)
In [4]	{1,14}, {2,11}, {3,10},{4,12},{5,8}, {6, 7},{9,13},{15}	73%	835
In[5]	{1,14}, {2,11,9,13},{3, 4,12},{5,8}, {6},{7}, {14} ,{15}	40%	623
In[6]	{1,14, 4,12,3,10},{2,11}, {5,8},{6},{7}, {9,13} ,{15}	60%	728
In[12]	{1},{2,11},{3,10},{4,12},{5,8},{6},{7},{9,13},{14},{15}	100%	432
Our algorithm	{1}, {2,11},{3,10},{4,12},{5,8},{6},{7},{9,13},{14},{15}	100%	213

As can be seen from the table, Our algorithm and the algorithm in [12] can get the correct result. In fact, the algorithm proposed in [12] is based on mean curvature co-occurrence matrix. However, our algorithm cost only 213 ms in the experiment because the plane points are discarded before matching. Therefore, the idea based on Gaussian curvature co-occurrence matrix can get better classification result to objects with slight different shape characteristic.

In the second experiment, firstly, we add Gaussian noise $(N(0,\sigma))$ on each object in the database. In this experiment σ increases from 0 mm to 2.0 mm. We classify the various noisy surfaces using Gaussian curvature co-occurrence matrix and four traditional approaches. The classification rates are shown in Fig. 2 for various σ values.

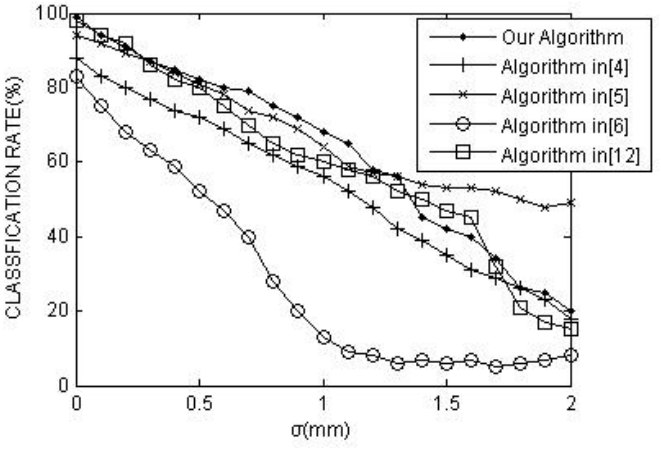


Fig. 2. Classification rate when Gaussian noise

We can see from the figure that our algorithm appears robust to Gaussian noise than some traditional approaches. But our algorithm has a lower classification rate than the method proposed in[5] when σ increases. However, our algorithm can keep a good classification rate when $\sigma < 1$ mm.

5 Conclusion

In this paper, Gaussian curvature is employed to represent the inherent characteristic of spatial curved surfaces, the conception of traditional co-occurrence matrix in texture analysis is extended to the idea of basing on Gaussian curvature of the spatial points, and then the Gaussian curvature co-occurrence matrix is constructed to describe the shape characteristic of 3D objects. Normalization method is proposed to form some invariants independent to the translation, scaling and rotation transform. Experiments indicate a better classification rate and running complexity than traditional approaches to 3D objects with slight different shape characteristic.

In the future work, the research of the recognition problems for occluded surfaces will be focused.

References

1. Huber, D., Kapuria, A., Donamukkala, R., Hebert, M.: Parts-based 3d object classification. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 82–89 (2004)
2. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271 (2004)
3. Wyngaerd, J., Gool, L., Koch, R., Proesmans, M.: Rotation invariant spherical harmonic representation of 3d shape descriptors. In: Proceedings of Symposium on Geometry Processing, Aachen, Germany, pp. 156–164 (2003)

4. Miao, Z., Gandelin, M., Baozong, Y.: Fourier transform based image shape analysis and its application to flower recognition. In: IEEE 6th International Conference on Signal Processing, Ottawa, Canada, vol. 2, pp. 1087–1090 (2002)
5. Novotni, M., Klein, R.: Shape retrieval using 3D Zernike descriptors. Computer-Aided Design 36(11), 1047–1062 (2004)
6. Laskov, P., Kambhamettu, C.: Curvature-based algorithms for nonrigid motion and correspondence estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), 1349–1354 (2003)
7. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Transactions on Graphics 21(4), 807–832 (2002)
8. Zhang, D., Hebert, M.: Harmonic maps and their applications in surface matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA, pp. 2524–2530 (1999)
9. Gupta, L., Sayeh, M.R., Tammana, R.: Neural network approach to robust shape classification. Pattern Recognition 23(6), 563–568 (1990)
10. Dubrovin, B., Fomenko, A., Novikov, S.: Modern geometry-methods and applications (I), 2nd edn. GTM, Springer, Heidelberg (1999)
11. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics 3(9), 610–621 (1973)
12. Guo, K., Liu, C., Yang, J.: 3D Objects Recognition Using Curvature Co-occurrence Matrix. Computer Science 35(7), 151–152 (2008) (in Chinese)
13. Besl, P.J.: Surfaces in range image understanding. Springer, New York (1988)
14. Besl, P.J., Jain, R.C.: Segmentation through variable-order surface fitting. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(2), 167–192 (1988)
15. Argenti, F., Alparone, L., Benelli, G.: Fast algorithms for texture analysis using co-occurrence matrices. IEEE Proceedings, Part F: Radar and Signal Processing 137(6), 443–448 (1990)

Nonlinear System Identification of Bicycle Robot Based on Adaptive Neural Fuzzy Inference System

Xiuli Yu, Shimin Wei, and Lei Guo

Beijing University of Posts and Telecommunications
Beijing, China
alicexiuli@163.com, wsmly@bupt.edu.cn, guolei@bupt.edu.cn

Abstract. System identification is the basis of designing control system. The bicycle robot is an under-actuated, non-linear, non-integrated system with lateral instability, it's two wheels are longitudinal and has non-sliding contact with the ground, meanwhile it's dynamic characteristics are complicated. So it is very difficult to set up more precise dynamics model. While precise model of complex system often requires more complex control design and calculation. In this paper, linear ARX model and nonlinear ANFIS model are proposed. The identifications of bicycle robot system are completed through the data of handlebar angle and those of inclination angle which are gathered when bicycle robot is stable. Simulation result by ANFIS based on T-S model could be very similar to the actual test data of bicycle robot sysytem, and it's identification precision is higher than that of linear ARX model. The obtained conclusions of fuzzy inference between input and output by above identificaton methods can provide some reference value for effective control on bicycle robot system in future.

Keywords: Bicycle robot; System identification; ARX Parameter model; ANFIS(Adpative-Network-based Fuzzy Inference Systems).

1 Introduction

The dynamic characteristics of Bicycle robot often show non-linear, time-varying, large delay, large inertia and other characteristics, which makes it difficult to establish a relatively accurate models of dynamatics. Mathematical model of the system is usually established in two ways, one is derived from the basic laws of physics models for the mechanism model, and the other one is running from the system and experimental data to establish a system identification model. For most complex systems, it is difficult to establish a relatively accurate models by the modeling method of mechanism [1]. System identification is the process of constructing a model to predict the behavior of a target system, and it is to find a equal model from a group of given models using the data of input and output, which points out the three major elements of system identification---the data of input and output, the model class and equal principle [2]. Conventional system identification techniques are mostly based on linear models with fast computation and rigorous mathematical support. In recent years, fuzzy neural network modeling represents nonlinear identification

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 381–388, 2010.

techniques that require massive computation but without mathematical proofs of convergence to global minima.The nature of system identification is an optimization problem, and to select the optimization standard depends on the goals of identification and the complexities of identification algorithm.This paper applies two representative approaches(ANFIS and ARX) from both disciplines and compare their performance on a classic system identification problem of bicycle robot[3].

This paper is organized into five sections. In the next section, the basics of ANFIS are briefly introduced. Section 3 explains the problem of bicycle robot modeling and how to use the ARX model to find a linear model. Section 4 exhibits the use of ANFIS for the same problem. Section 5 gives concluding remarks.

2 Basic Theory of ANFIS

Adaptive neural fuzzy inference system(ANFIS) is a product by combining the fuzzy inference system with the neural networks. ANFIS fully makes use of the excellent characteristics of the neural network and the fuzzy inference system, and is widely applied in many fields of fuzzy controller design and model identification. As a special neural network, ANFIS can approximate all nonlinear systems with less training data and quicker learning speed and higher precision. ANFIS is a neural network in fact, which realize sugeno fuzzy system using network.

Jang's ANFIS [4] is a 5-layered feedforward neuro-fuzzy network whose node functions of the same layer are of the same function family. ANFIS applies the rules of TSK (Takagi, Sugeno and Kang) form in its architecture. Let us consider a physical system with m input variables $x = [x_1, x_2, ..., x_m]^T$ and a single output y, with

$x_i \in D_i = [x_{i,\min,} x_{i,\max}] \subseteq R \quad (i = 1, .., m)$

and $y \in Y = [y_{\min,} y_{\max}] \subseteq R$. The TSK fuzzy rules are of the following form:

R(r):

IF x_1 is $A^r{}_1$ AND ... AND x_m is $A^r{}_m$ THEN

$$y_r = g^r(x) = \alpha^r{}_0 + \alpha^r{}_1 x_1 + ... + \alpha^r{}_m x_m \quad (r = 1, ..., n). \tag{1}$$

where n is the number of rules, x_i ($i = 1, .., m$) are the model inputs, and y_r is the output of the rule $R^{(r)}$. $A^r{}_i$ ($i = 1, .., m$) are fuzzy sets in D_i characterized by membership functions $\mu_{A^r{}_i}(x_i)$ and $\alpha^r{}_i$ are real coefficients of the rule polynomials, $g^r(x)$. the m-tuple ($A^r{}_i$, ..., $A^r{}_m$) of fuzzy sets in the premise part of $R^{(r)}$ forms a fuzzy region $A^r{}_i \times ... \times A^r{}_m$ in $D = D_1 \times ... \times D_m$, called the fuzzy hyper-cube (FHC) in this paper.

Considering a first-order TSK fuzzy inference system which contains the following rules in identification of bicycle robot system:

Rule r: If x_1 is $A^1{}_i$ and x_2 is $A^2{}_j$ and x_3 is $A^3{}_k$, then $y_r = s^r + p^r x_1 + q^r x_2 + r^r x_3$,where $r \in \{1,...,27\}$ and $i, j, k \in \{1,2,3\}$,

The corresponding ANFIS architecture is depicted in Fig. 1. Circles in ANFIS represent fixed nodes.Fixed nodes function as predefined operators to their inputs and no other parameters but the inputs participate in their calculations. Adaptive nodes, on the other hand, have some internal parameters which affect the results of their calculations[5].

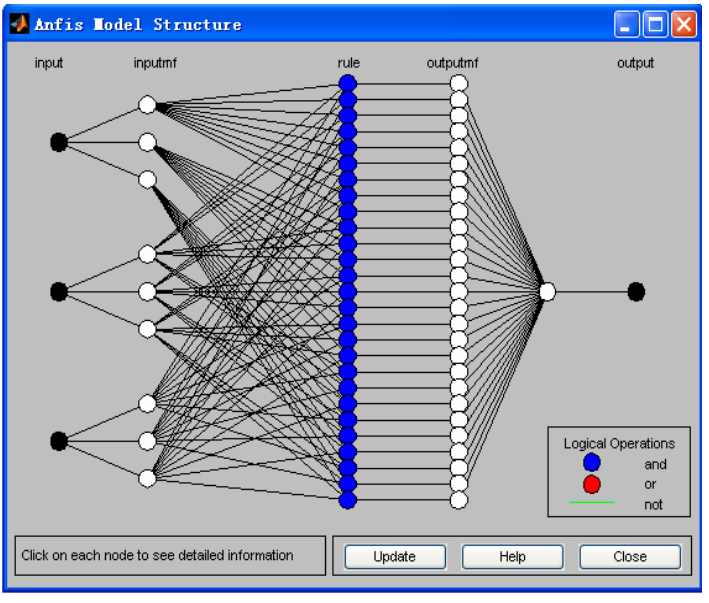


Fig. 1. An adaptive-network-based fuzzy inference system

Layer 1 is an adaptive layer which denotes membership functions to each input. In this paper we choose Gaussian functions as membership functions[6, 7]:

$$O^1_i = \mu_{A_i}(x) = \exp^{-(\frac{x-c_i}{a_i})^2}. \tag{2}$$

where x is the input to node i; Ai is the membership function associated with this node; and $\{c_i, a_i\}$ is the parameter set that changes the shapes of the membership function. Parameters in this layer are referred to as the premise parameters.

Layer 2 is a fixed layer in which each node calculates the firing strength of each rule via multiplication:

$$O^2_i = w_i = \mu_{A_i}(x_1) \times \mu_{B_i}(x_2) \times \mu_{C_i}(x_3)., \quad \text{i=1,2,3} \tag{3}$$

Layer 3 is a fixed layer in which the i-th node calculates the ratio of the i-th rule's firing strength to thesum of all rule's firing strength:

$$O^3_j = \overline{w_j} = \frac{w_j}{\sum_{i=1}^{27} w_i}, \text{ j = 1,2,...27.} \tag{4}$$

Layer 4 is an adaptive layer in which the i-th node deals with the consequent parameters of the i-th rule as mentioned in (1). Node i in this layer has the following node function:

$$O^4{}_j = \overline{w_j} f_j = \overline{w_j}(p^i x_1 + q^i x_2 + r^i x_3 + s^i),\ \text{i=1,2,...27} \tag{5}$$

where w- is the output of i-th node of layer 3 and $\{\alpha^i{}_1, \alpha^i{}_2, \alpha^i{}_3\}$ is the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5 is a fixed single-node layer which computes the overall output as the summation of all incoming signals:

$$O^5{}_j = \sum_{j=1}^{27} \overline{w_j} f_j = \frac{\sum_{j=1}^{27} w_j f_j}{\sum_{i=1}^{27} w_i} \tag{6}$$

ANFIS is a special neural network, if input variables are divided into enough fuzzy sets, the network can accurately approximate all kinds of nonlinear function by adjusting parameter of membership function in the first layer and adjusting output function parameter in the fourth layer.

3 Bicycle Robot Modeling and ARX Model

In order to analyze the relationship between handlebar angle and the inclination angle of bicycle during bicycle robot runing stably, linear and nonlinear identification have been done based on the ARX model and adaptive network-based fuzzy inference systems(ANFIS) by using the test data of tilt angel and handlebar angle. To proceed system identification on bicycle robot, firstly,we need to collect input-output data pairs through the ARM9 Embedded Processor. The sampling time is 0.03 second. The datas are consisted of two parts: the input $u(k)$ is the voltage of handlebar rotation angle, which is potentiometer output. And the output $y(k)$ is the tilt angle β of bicycle robot. Tilt angle β through three-dimensional attitude measurement system MTi measured. As the three-dimensional attitude measurement system built-in three-axis gyroscopes and three-axis acceleration, real-time data updates, accesses to high-precision values of the body tilt angle.

In this paper, From Fig. 2, one thousand input-output data points were collected at 1.1m / s speed of the bicycle robot for the system identification model.

ARX model is to remove the means from the data and assume a linear model of the form:

$$y(k)+a_1 y(k-1)+\cdots+a_{n_a} y(k-n_a)=b_1 u(k-n_k)+\cdots+b_{n_b} u(k-n_k-n_b+1) \tag{7}$$

n_a is the rank scopes of the polynomial output of ARX model, n_b is the rank scopes of the polynomial input of ARX model, n_k is the scopes of time extended[8] .

To find an ARX model for bicycle robot, the data set was divided into training (k = 1 to 300) and test (k = 301 to 600) data sets. We performed an exhaustive search on

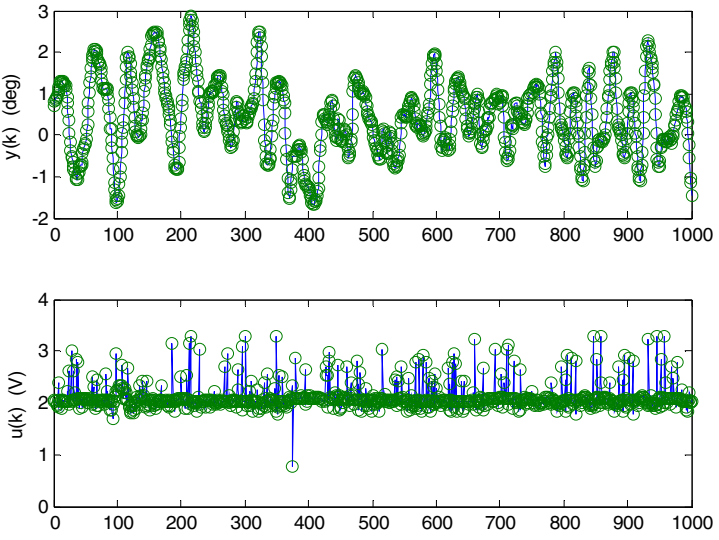


Fig. 2. 1.1m/s input-output data pairs

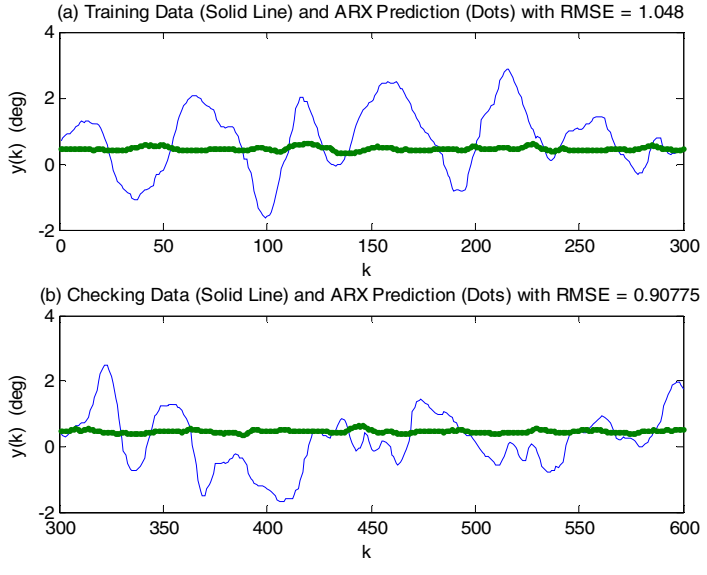


Fig. 3. The RMSE comparison of training data, checking data and prediction data of ARX

the ARX structure $[n_a, n_b, n_k]$, where each of the structure parameters is allowed to changed from 1 to 10 independently. In other words, we constructed 1000 ARX models whose parameters are identified by least-squares methods. The best model was selected as the one with the smallest test error. and the best ARX model is

$[n_a, n_b, n_k]$ = [2,5,10], with training RMSE 1.048 (root-mean-squared error) and test RMSE 0.90775. Figure 3 demonstrates the fitting results of the ARX model. However, if a better performance is desired, then we have to resort to nonlinear models.

4 ANFIS Model of Bicycle Robot

To use ANFIS for system identification, the first thing we need to do is input selection, i.e., to determine which variables should be the input arguments to the ANFIS model. Once the input arguments are fixed, then we can specify the ANFIS model structure, such as the style for input space partitioning, the numbers and types of membership functions on each input, and so on.

For the bicycle robot modeling problem, we can partition the input candidates ito two disjoint sets:

$$Y = \{y(k-1), y(k-2), y(k-3), y(k-4),\}$$

$U = \{u(k-1), u(k-2), u(k-3), u(k-4), u(k-5), u(k-6),\}$ The exhaustive search approach for input selection is used, and non-linear model identification of bicycle robot can be completed through the adaptive neuro-fuzzy inference system (ANFIS).

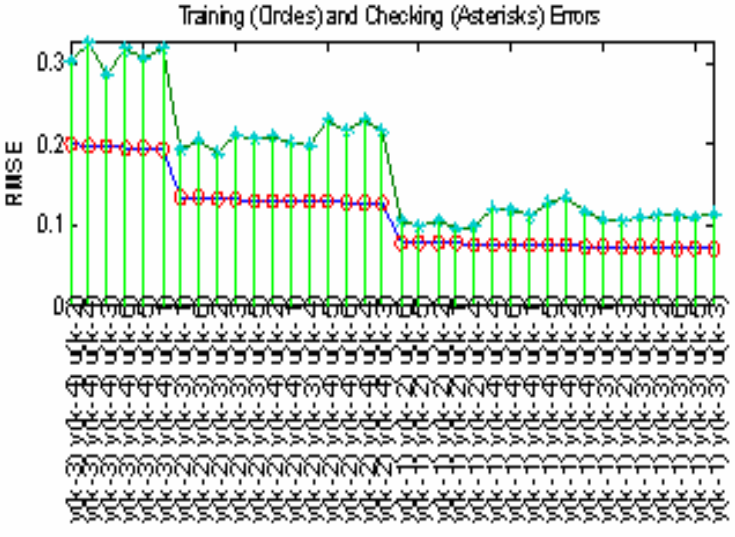


Fig. 4. Training RMSE and checking RMSE of ANFIS

From Fig. 4, respectively, it can be seen that the system finally selected three input $y(k-1), y(k-3),$ and $u(k-3)$ at the speed of 1.1m/s. Fig.5 shows the simulation results of ANFIS approaches, and its training RMSE = 0.068796, checking RMSE = 0.11223.

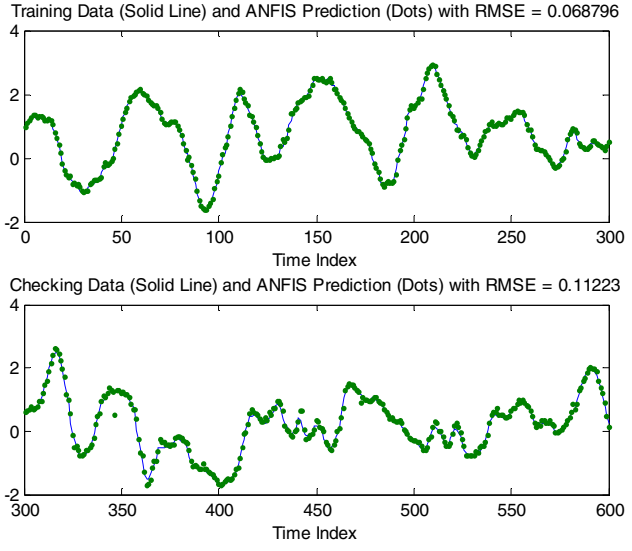


Fig. 5. The RMSE comparison of training data, checking data and prediction data of ANFIS

5 Concluding Remarks

This paper mainly studies the bicycle robot system identification on relation between handlebar angle and inclination angle of bicycle through the data of handlebar angle and those of inclination angle which are gathered when bicycle robot is stable. We have compared the performance of bicycle robot modeling by using two approaches: ANFIS and ARX model. It is obvious that the RMSE and error of data is less using ANFIS than ARX model. We can draw a conclusion that Simulation result by ANFIS based on T-S model could be very similar to the actual test data of nonlinear bicycle robot sysytem, and it's identification precision is higher than that of linear ARX model. The obtained conclusions of fuzzy inference between input and output by above identificaton methods can provide some reference value for effective control on bicycle robot system in future.

References

1. Fu-Cai, L.: Fuzzy Model Identification and Application of Nonlinear Systems. National Defense Industry Press, Beijing (2006)
2. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its application to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics 51(1), 116–132 (1985)
3. Xiu-ying, L., Zhi-gang, H.: Advances in Nonlinear System Identification. Techniques of Automation & Application 23(4), 5–7 (2004)
4. Zhi-Xiang, H., He-Qing, L.: Nonlinear System Identification Based on Adaptive Neural Fuzzy Inference System. In: 2006 International Conference on Communications, Circuits and Systems Proceedings, vol. 3, pp. 2067–2069 (2006)

5. Jang, J.-S.: ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Trans. Syst., Man., Cybern. 15, 116–132 (1985)
6. Rastegar, F., Araabi, B.N., Lucast, C.: An evolutionary fuzzy modeling approach for ANFIS architecture. In: The 2005 IEEE Congress on Evolutionary Computation, vol. 3, pp. 2182–2189 (2005)
7. Yue, J., Liu, J., Liu, X., Tan, W.: Identification of nonlinear system based on ANFIS with subtractive clustering. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, pp. 1852–1856 (2006)
8. Li-Wei, D., Zhi-Hua, W., Zhi-Wei, X.: Identification Research of the Piezoelectric Smart Structure System of Aircraft Wing Based on ARX Model. Piezoelectrics & Acoustooptics 30(6), 760–762 (2008)

Transmission: A New Feature for Computer Vision Based Smoke Detection

Chengjiang Long, Jianhui Zhao*, Shizhong Han, Lu Xiong,
Zhiyong Yuan, Jing Huang, and Weiwei Gao

Computer School, Wuhan University, Wuhan, Hubei, 430072, PR China
jianhuizhao@whu.edu.cn

Abstract. A novel and effective approach is proposed in this paper to detect smoke using transmission from image or video frame. Inspired by the airlight-albedo ambiguity model, we introduce the concept of transmission as a new essential feature of smoke, which is employed to detect the smoke and also determine its corresponding thickness distribution. First, we define an optical model for smoke based on the airlight-albedo ambiguity model. Second, we estimate the preliminary smoke transmission using dark channel prior and then refine the result through soft matting algorithm. Finally, we use transmission to detect smoke region by thresholding and obtain detailed information about the distribution of smoke thickness through mapping transmissions of the smoke region into a gray image. Our method has been tested on real images with smoke. Compared with the existing methods, experimental results have proved the better efficiency of transmission in smoke detection.

Keywords: smoke detection, dark channel prior, soft matting; transmission.

1 Introduction

Traditional smoke and fire detections which have been widely applied in the buildings are based on sensors. Such detection approaches require a very close proximity to fire or smoke and are often disturbed by a variety of noises, so they may be not reliable and cannot be spread into open spaces and larger areas. Vision based fire detection approaches make it possible to serve large and open spaces, such as auditoriums, tunnels and forest, and they also provide abundant visual information about fire or smoke. Within the computer vision based fire detection, smoke location and analysis are very important, especially in the cases that flame is covered by smoke during burning or there is obvious smoke but little flame during the initial procedure of fire disaster.

Color, shape and texture are three usually used important features in detecting smoke from one single image or the video sequence. Color is widely used in either single image or multiple frames, while shape and texture are mostly utilized in continuous video frames. Color feature provides a clue as the precondition to locate the possible smoke regions. Chen et al. [1] and Yuan [2] proposed chrominance

* Corresponding author.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 389–396, 2010.

detection approaches which specify the color ranges of smoke in RGB color space and the intensity ranges of smoke in HSI color space. Shape feature is used as a good indication of the presence of smoke in the view field of camera, and its variation is analyzed by geometry differences of successive frames, which can be implemented by background subtraction or neighbor frames subtraction. Fujiwara [3] assumed that smoke shapes have the property of self-similarity, and thus proposed the technique for extracting smoke regions from an image using fractal encoding concepts. Chen et al. [4] used a shape disorder decision rule involving the quantity of smoke pixels to verify a real smoke extracted by the color feature from the captured image sequences. Xu et al. [5] extracted features of the moving target including growth, disorder, and self-similarity as important components of a joint shape feature vector to recognize fire smoke. Toreyin et al. [6] detected the boundary of smoke region and then used the high frequency nature of boundaries represented in wavelet domain as a clue to model the smoke flicker, which helps detect smoke in video. Texture is another feature which can express the difference between smoke and non-smoke, and is usually extracted and analyzed by two popular approaches including Gray Level Co-occurrence Matrices (GLCM) [7] and wavelet-based analysis [8]. Cui et al. [9] investigated the smoke texture and the non-smoke texture with wavelet packet and GLCM, and then a neural network is used to help judge whether a fire occurs. Fazekas et al. [10] proposed a method for detecting regions with dynamic texture in video sequences, and segmented the frame into static regions and dynamic regions using a level set scheme. Yu et al. [11] proposed a method of texture analysis for real-time fire smoke detection, using neural network to classify smoke and non-smoke regions based on the texture features. Ferari [12] adopted the discrete wavelet transformation on suspicious regions, describe the intensity of texture features with wavelet coefficients and then determine whether smoke exists according to whether the details of texture features reduce.

However, the existing methods using the aforementioned smoke features cannot achieve satisfying results: (1) they can locate heavy smoke regions but have difficulty to detect the light smoke regions; (2) they cannot directly provide the thickness distribution of smoke region which is important for fire estimation. Therefore, it is necessary to investigate new and more essential features for smoke to solve these two problems in smoke detection.

In recent years, some research works are carried out on haze removing based on the airlight-albedo ambiguity model. Inspired by the model, we introduce the concept of transmission into smoke detection and test its validity through experiments. In haze removal algorithms [13-14], transmission is only used to recover the scene radiance of the image. Comparatively, we use transmission as a new feature of smoke to detect smoke region and its thickness distribution. Based on our knowledge, there is no such attempt previously.

The paper is organized as follows. Section Ⅱ presents the optical model for smoke. In Section Ⅲ, smoke transmission is preliminarily estimated using dark channel prior, then refined through soft matting algorithm, and then smoke regions can be detected and their thickness distributions can be described based on the calculated transmissions. Section Ⅳ provides the experimental results on real images. Section Ⅴ concludes this paper with brief discussion.

2 Optical Model for Smoke

Based on References [13-17], the final image observed intensity consists of two parts: scene radiance and global atmospheric light, while the relationship between these two parts can be represented with the airlight-albedo ambiguity model described as:

$$I(x) = \tau(x)J(x) + (1-\tau(x)A \tag{1}$$

For one pixel $x(i, j)$ of the image with three RGB color channels, $I(x)$ stands for the observed intensity, A is the global atmospheric light, $J(x)$ is the surface radiance vector at the intersection point of the scene and the real-world ray corresponding to pixel x, and $\tau(x)$ describes the portion of the light that is not scattered and reaches the camera.

Considering that in our work we focus on smoke detection and take transmission as the essential feature of smoke, we specially define an optical model based on Equation (1) for smoke as:

$$I(x) = (1-t(x))J(x) + t(x)S \tag{2}$$

In our defined model, S is the global smoke color vector which is constant and available to reference, and $t(x)$ is the smoke transmission of pixel x describing the portion of S that contributes to the ultimate observed intensity.

Therefore, the defined model of Equation (2) can be applied for the image in the presence of smoke. By calculating the smoke transmission $t(x)$ we can judge whether the pixel x belongs to the smoke region or not. If pixel x is decided to be within the smoke region, transmission $t(x)$ can imply the information about smoke thickness at the pixel.

3 Estimate the Smoke Transmission

3.1 Preliminary Estimation Using Dark Channel Prior

The dark channel prior is described as: at least one channel among R, G and B color channels has very low intensity at some pixels in a block. According to Reference [14], the color channel whose value is the lowest is called dark channel. In most cases of the real world, color of smoke may be prone to white color and the dark channels in the areas where smoke exists have high values. Therefore, the intensity of the dark channel can be taken as a rough approximation to describe the thickness of smoke, and we use the intensity property to estimate the smoke transmission preliminarily.

Based on the descriptions above, we express the dark channel for an image J as:

$$J_{dark}(x) = \min_{c\in\{r,g,b\}}\left(\min_{y\in\Omega(x)}\left(J_c(y)\right)\right) \tag{3}$$

where $\Omega(x)$ is the local block or window with pixel x as the center point, and the block size can be adjusted with parameter.

We denote $\tilde{t}(x)$ as smoke transmission of block $\Omega(x)$, which is assumed as the minimum value of transmission values of all the pixels in region $\Omega(x)$. We take $\tilde{t}(x)$ as the preliminary transmission of any pixel in $\Omega(x)$, i.e. all pixels in block $\Omega(x)$ have the same preliminary transmission. Then the min operation is performed for $\Omega(x)$, and Equation (2) can be rewritten as:

$$\min_{y\in\Omega(x)}\left(I_c(y)\right)=\left(1-\tilde{t}(x)\right)\min_{y\in\Omega(x)}\left(J_c(y)\right)+\tilde{t}(x)S_c \tag{4}$$

Considering that S_c is the constant global smoke color vector and is always positive, after both sides are divided by S_c, Equation (4) is equivalent to:

$$\min_{y\in\Omega(x)}\left(\frac{I_c(y)}{S_c}\right)=\left(1-\tilde{t}(x)\right)\min_{y\in\Omega(x)}\left(\frac{J_c(y)}{S_c}\right)+\tilde{t}(x) \tag{5}$$

Notice that the min operation is performed on three color channels independently, we still need to choose the minimum value among these three values on Equation (5). Thus we have:

$$\min_{c}\left(\min_{y\in\Omega(x)}\left(\frac{I_c(y)}{S_c}\right)\right)=\left(1-\tilde{t}(x)\right)\min_{c}\left(\min_{y\in\Omega(x)}\left(\frac{J_c(y)}{S_c}\right)\right)+\tilde{t}(x) \tag{6}$$

According to the conclusion from the dark channel prior, the dark channel J_{dark} of the smoke-free radiance J should be prone to zero:

$$J_{dark}(x)=\min_{c\in\{r,g,b\}}\left(\min_{y\in\Omega(x)}\left(J_c(y)\right)\right)=0 \tag{7}$$

Obviously,

$$\min_{c\in\{r,g,b\}}\left(\min_{y\in\Omega(x)}\left(\frac{J_c(y)}{S_c}\right)\right)=0 \tag{8}$$

By substituting Equation (8) into Equation (6), the block's smoke transmission (or the preliminary transmission of any pixel in the region) $\tilde{t}(x)$ can be derived as following:

$$\tilde{t}(x)=\min_{c}\left(\min_{y\in\Omega(x)}\left(\frac{I_c(y)}{S_c}\right)\right) \tag{9}$$

3.2 Refine Preliminary Estimation through Soft Matting

Although the smoke transmission $t(x)$ of one pixel can be estimated simply and approximately with $\tilde{t}(x)$ depending on foregoing assumptions, the equations (2)-(8) above are based on blocks rather than specific pixel. To obtain the more precise transmission for each pixel, the soft matting algorithm [18] is used to refine the preliminary transmission $\tilde{t}(x)$. Based on Reference [18], the optimal transmission can be acquired through solving the following large sparse linear equation:

$$(L+\lambda U)t = \lambda\tilde{t} \tag{10}$$

In Equation (10), t and $\tilde{t}$ are denoted as the vector form of $t(x)$ and $\tilde{t}(x)$ respectively, L is the Matting Laplacian matrix put forward by Levin [18], λ is a regularization parameter and we set the value as 10^{-4} in our experiment, U is an identity matrix of the same size as L. Any element in matrix L can be described as:

$$L(i,j)=\sum_{k|(i,j)\in w_k}\left(\delta_{ij}-\frac{1}{|w_k|}\left(1+(I_i-\mu_k)^T\left(\Sigma_k+\frac{\varepsilon}{|w_k|}U_3\right)^{-1}(I_j-\mu_k)\right)\right) \tag{11}$$

In Equation (11), I_i and I_j are the color vectors of the ith pixel and the jth pixel of the input Image I, δ_{ij} is the Kronecker delta, μ_k and Σ_k are the mean and covariance matrix of the color vectors in window w_k, U_3 is a 3×3 identity matrix, ε is a regularizing parameter, and $|w_k|$ is the number of pixels in the window w_k.

To solve the large sparse linear Equation (10), we choose the Preconditioned Conjugate Gradient (PCG) algorithm as solver. By solving this equation, the outcome of transmission t for each pixel can be used to detect the smoke regions and imply the corresponding distribution of smoke thickness.

3.3 Smoke Detection Based on Transmission

For images used in experiment, we set the value of S_c as the color vector of the pixel whose dark channel value is the largest in the related image. As for the rule of smoke detection based on transmission, we consider a pixel as a smoke pixel belonging to smoke region if its value of smoke transmission is larger than the threshold value, otherwise it is taken as a non-smoke pixel. We denote the threshold value and the maximum value of smoke transmission as t_0 and $t_{\max}$ respectively. To represent the thickness distribution of smoke region, we simply map the smoke transmission range $[t_0, t_{\max}]$ to the gray color range $[t_0\times255, 255]$ by:

$$G(x) = \begin{cases} \frac{(t(x)-t_0)}{(t_{\max}-t_0)} \times (255 - t_0 \times 255) & if\ t(x) > t_0 \\ 0 & else \end{cases} \tag{12}$$

From Equation (12), we can find that the grayish value is larger where the smoke transmission is larger, therefore the gray image can visually imply the smoke thickness distribution in the detected smoke region.

4 Experimental Results

The proposed novel approach for smoke detection is implemented on PC with 2.79 GHz Intel(R) Core(TM) 2 Duo CPU and 2.0 GB RAM using OpenCV library. To validate the performance of our approach, the existing method based on color feature from Reference [2] is used for comparison.

Experimental results on real images with heavy smoke are shown in Figure 1. It can be found that our approach detects the smoke regions accurately, and also provides more detailed information about thickness distribution, i.e. smoke is thicker where the mapped gray value from transmission is larger.

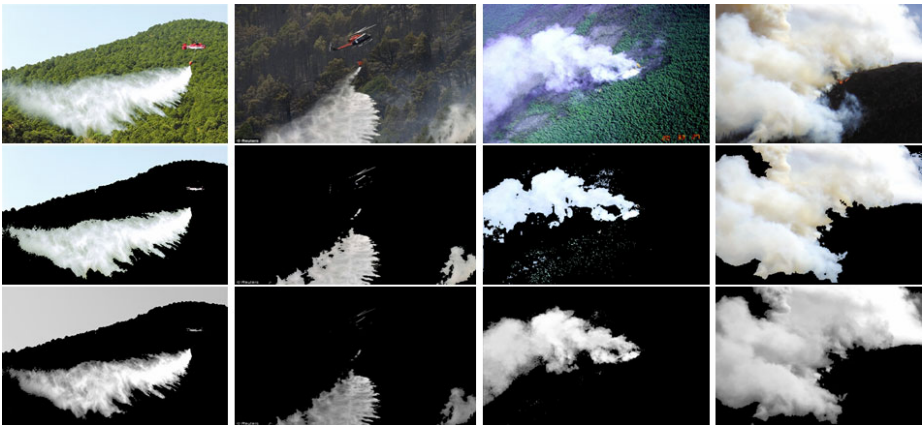


Fig. 1. Detection results of heavy smoke, Top: original images; Middle: results from Reference [2]; Bottom: our results

Experimental results on real images with light smoke are shown in Figure 2. The semi-transparent light smoke often exists with fire, and it can appear with non-grayish color caused by the covered objects, e.g. soil, trees, or flame. Thus the existing methods using color, shape or texture features have difficulties to detect the light smoke. Different with them, we adopt transmission as one essential smoke feature, and can greatly reduce the effects from covered objects.

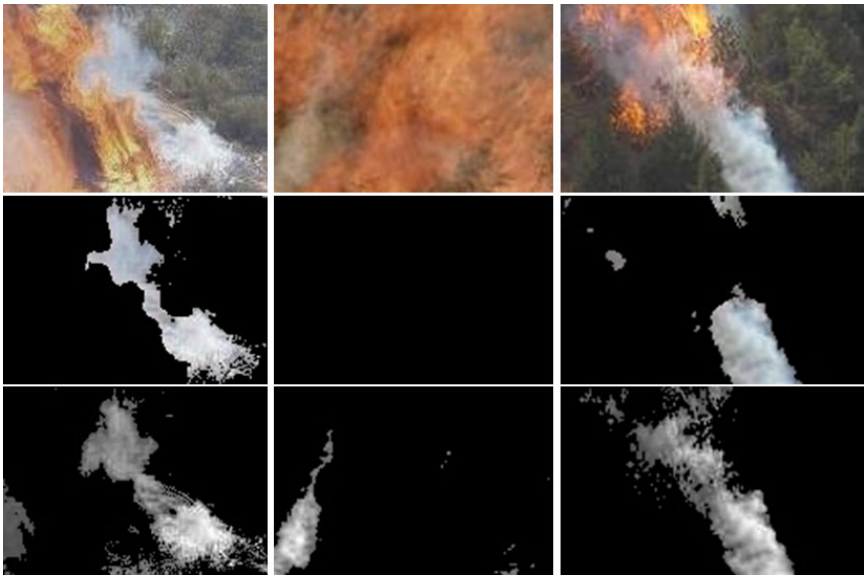


Fig. 2. Detection results of light smoke, Top: original images; Middle: results from Reference [2]; Bottom: our results

5 Conclusion

In this paper, a novel approach to detect smoke using transmission is proposed. We estimate the preliminary smoke transmission using dark channel prior and then refine the result through soft matting algorithm. According to the calculated transmission value, smoke pixels can be detected accurately and the detailed information about thickness distribution of smoke region can also be provided directly. Our method has been tested on real images with heavy or light smoke, and the experimental results have proved its efficiency.

Of course, the estimation method of smoke transmission is preliminarily based on dark channel prior, so our approach is currently limited in detecting gray-white smoke. Although most smoke in natural fire disasters is prone to gray-white, in a few cases there really exist smoke of other colors and gray-white objects like motionless marbles, which will affect the accuracy of our method. In the future, we will add color range to recognize smoke with other colors and take motion feature to eliminate the effect from gray-white non-smoke objects.

Acknowledgments. This work was supported by Hubei Provincial Natural Science Foundation of China, Research Foundation (No. AISTC2008_16) from State Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, Fundamental Research Funds for the Central Universities, and the 985 Project of Cognitive and Neural Information Science, Wuhan University (No. 904273258).

References

1. Chen, T.H., Yin, Y.H., Huang, S.F., Ye, Y.T.: The Smoke Detection for Early Fire-Alarming System Base on Video Processing. In: Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (2006)
2. Yuan, F.N.: A fast accumulative motion orientation model based on integral image for image for video smoke detection. J. Pattern Recognition Letters 29, 925–932 (2008)
3. Fujiwara, N., Terada, K.: Extraction of a Smoke Region Using Fractal Coding. In: International Symposium on Communications and Information Technologies, Japan, pp. 659–662 (2004)
4. Chen, T.H., Wu, P.H., Chiou, Y.C.: An Early Fire-Detection Method Based on Image Processing. In: IEEE International Conference on Image Processing, Singapore, pp. 1707–1710 (2004)
5. Xu, Z.G., Xu, J.L.: Automatic Fire Smoke Detection Based on Image Visual Features. In: International Conference on Computational Intelligence and Security Workshops, pp. 316–319 (2007)
6. Toreyin, B.U., Dedeoglu, Y., Cetin, A.E.: Contour Based Smoke Detectio In Video Using Wavelet. In: 14th European Signal Processing Conference EUSIPCO, Florance (2006)
7. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, 610–621 (1973)
8. Kara, B., Watsuji, N.: Using Wavelets for Texture Classification. J. WSEAS Transactions on Computers, 920–924 (2003)
9. Cui, Y., Dong, H., Zhou, E.Z.: An Early Fire Detection Method Based on Smoke Texture Analysis and Discrimination. J. Congress on Image and Signal Processing, 95–99 (2008)
10. Fazekas, S., Amiaz, T., Chetverikov, D., Kiryati, N.: Dynamic Texture Detection Based on Motion Analysis. J. Int. J. Comput. Vis. 82, 48–63 (2009)
11. Yu, C.Y., Zhang, Y.M., Fang, J., Wang, J.J.: Texture Analysis of Smoke For Real-time Fire Detection. In: Second International Workshop on Computer Science and Engineering, pp. 511–515 (2009)
12. Ferari, R.J., Zhang, H., Kube, C.R.: Real-time Detection of steam in video images. J. Pattern Recognition 40, 1148–1159 (2007)
13. Fattal, R.: Single image dehazing. In: SIGGRAPH, pp. 1–9 (2008)
14. He, K.M., Sun, J., Tang, X.O.: Single Image Haze Removal Using Dark Channel Prior. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2009)
15. Tan, R.T.: Visibility in bad weather from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008)
16. Narasimhan, S.G., Nayar, S.K.: Chromatic framework for vision in bad weather. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 598–605 (2000)
17. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. J. IJCV 48, 233–254 (2002)
18. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 61–68 (2006)

A Novel Features Design Method for Cat Head Detection

Hua Bo

School of Information Enginnering, Shanghai Maritime University,
201303, Shanghai, China
huabo@shmtu.edu.cn

Abstract. In this paper we have proposed a new novel features model whichdesigned to robustly detect the highly variable cat head patterns.Do not like human, cats usually have distinct different face, pose,appearance and different scales of ears, eyes and mouth. So manysignificant features on human face detection have presented but itis not satisfying to use them on cat head. We have designed a newfeatures model by ideally combining the histogram frame withGLCM-based (gray level co-occurrence matrix) texture features todescribe both the shape information of cat's head and textureinformation of cat's eyes, ears and mouth in detail. SVM-basedclassifier achieves the detection results. Extensive experimentalresults illustrating the high detection rate with low false alarm.

Keywords: Cat Head Detection, Gray Level Co-occurrence Matrix, Support Vector Machine.

1 Introduction

This paper designed a novel feature model for detecting the head ofcat-like animals in images, adopting cat as a test case. Petdetection or cat detection has a few works now, but the similarproject, for example, human detection has been developed in recentyears.

Significant research works of human detection include: Human [1][2] and face detection [3][4], achievingnear 90% detection rate on the frontal face in real-time [4]using a boosting based approach. Local face features, two class'sfeatures, image features and gradient features, which belong to thelow level features of images, play a crucial role in objectsdetection. The image features are directly extracted from theimages, such as intensity values[3], imagepatch[5], PCA coefficients[6], and waveletcoefficients[4][7][8]. The image featuresare suitable for small window and usually require a good photometricnormalization. Contrarily, the gradient features are more robust toillumination changes. The gradient features are extracted from theedge map[9][10] or oriented gradients, which mainlyinclude SIFT[11], EOH[12], HOG[1]covariance matrix[13], shapelet[14], andedgelet[15]. The literatures[1][17][18] showthat locally normalized Histogram of Oriented Gradient (HOG)descriptor provide excellent performance relative to other existingfeature sets including wavelet.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 397–405, 2010.

Human detection and cat head detection have the same applications onfilm and media analysis or visual surveillance. On these applicationbackgrounds the two works have the same problems as wide variety ofarticulated poses, variable appearance, complex backgrounds,unconstrained illumination and different scales,etc. But, on theother hand, they have important different aspect in detail analysis:human have relative coherent proportion of facial features and thefour limbs of the body, but cat head donot always has such nice andclear character. Figure 1 illustrates this idea.Figure 1 compares human body and cat head by addingand averaging about 2000 sample images of human body and cat head.Figure1(a) is the result of human body, it hasdistinct body profile. Figure1(b) is the result ofcat head aligned by ears, it shows ears' legible profile in thefirst half of image. Figure1(c) is the result of cathead aligned by eyes, it has vivid eyes' details. These figures showthat cat heads have more rich textures information than human body.

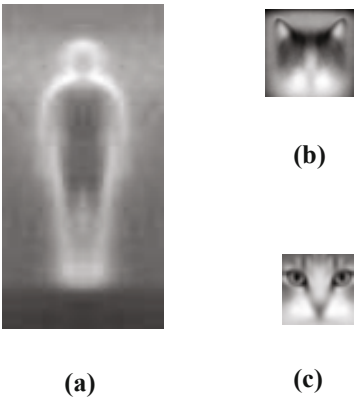


Fig. 1. Comparison of three average samples: (a) human body, (b) cat head aligned by ears, (c) cat head aligned by eyes

On such idea literature[16] authors proposed a joint shapeand texture detection method for cat heads by using the features ofin-channel HOOG and orthogonal-channel HOOG. On both shape detectorand texture detector using their own cat data set, compare of HOOGfeatures with conventional features as Haar, Haar+EOH and HOG arepresented also. The results are: 1) use shape detector, for giventotal false alarm count 200, the recall of Haar, Haar+EOH, HOG andthat paper proposed are 0.53/0.63/0.72/0.78. 2) Use texturedetector, for given total false alarm count 200, the recall of Haar,Haar+EOH, HOG and that paper proposed are 0.52/0.71/0.67/0.79. TheHaar feature using all four kinds of Haar templates on intensitygives the poorest performance because of large shape and texturevariations of the cat heads. With help of oriented gradientfeatures, Haar + EOG improve the performance. HOG features performbetter on the shape detector than on the texture detector. Usingin-channel and orthogonal-channel information, the detector based onHOOG features produce the best results. HOG and similar features aresensitive to edge information and they have good

performance inhuman detection. However, they maybe lost if the objects to detecthave no clear edge information.

Do not like human, cats usually have distinct different face, pose,appearance. Meanwhile their ears, eyes and mouth usually be indifferent size. So efficient details descriptions for cat's head isimportant. In this paper, we proposed a new feature model, whichideally combing the histogram frame with GLCM-based texture featureswhich describe the detailed information of eyes, ears and mouth ofthe cat's head. In this method space cells were composed by definiteorder at first. Some features were computed for every cell then andformed feature vectors which had the alike form as histogram. In theexperiments, we compare the performance of several features, eachcorresponding to a frontal upright face detector. Performanceevaluation has been carried out by a 10-fold cross validation andusing well labeled cat images downloaded fromhttp://mmlab.ie.cuhk.edu.hk/ and PASCAL 2007 cat data set.

This paper proposed a new feature design method which focused on the details texture description on some part of cat head. Section 2 describes the new feature and their algorithm forconstructing the feature vectors. Experimental results are shown inSection 3.

2 Our Proposed Feature Design Method

HOG and EOH are two classical features and HOG have good performancein human face detection. But HOG don't have very good performance indetecting cat heads. We compared our feature's false alarm andrecall values with HOG in later.

2.1 Histogram of Oriented Gradients (HOG)

HOG and EOH features have the same feature cell with different formvector. The basis unit in the HOG cell is the weighted orientationhistogram which is a small spatial region, e.g., 8*8 pixels.Usually it is represented as:

$$HOG\_cell(R) = [s^1(R), s^2(R),\ldots,s^k(R)]$$

Where K is the orientations sub images. Every cell S(R) expressesthe orientation gradients information.

But the HOG feature is not good for cat head detection because theorientation gradients information can't to express the rich textureinformation of cat head perfectly. Varying pose, Illuminationchange, partial occlusion, and different scale size of eye or ears,all of these will affect the orientation gradients uniformity.

2.2 Our Proposed Method: GLCM-Based Texture Histogram (HOT)

The cat head template is relatively rich in texture, features e.g.eyes, nose and mouth can carry enough discriminative information totell them apart from almost possible background images. We use thesetextures information by adding GLCM-based textures in featurevectors.

The basic unit of HOT descriptor is small regions or cells. Thesecells were obtained by slide window which pass through whole sampleimages step by step. All the cells are aligned in order and form ahistogram which is called GLCM-based texture histogram. HOT can berepresented in a same form as HOG instead oriented gradients withtextures information. It can be represented as:

$$\text{Feature\_cell}(R) = [H^1(R), H^2(R),\ldots,H^k(R)]$$

R means small region or cell, and K means the number of textures.The overlapped cells grouped and normalized to form a large specialregion called "block".

The textures are calculated form GLCM method. GLCM is an imagetexture analysis technique. The GLCM for an image is a matrix ofjoint probabilities $p_{i,j}(\theta,s)$with which two neighboringpixels separated by distance s at an angle ɵ occur on theimage, one with grey level i and the other with gray level j.Such matrices of gray spatial dependence frequencies are a functionof the angular relationship between the neighboring resolution cellsas well as a function of the distance between them.

It is difficult to analyze the texture or structure of an image byGLCM because its large matrix forms. So it is often to use theassociated texture features that are computed using only GLCMvalues. In literature[20]some features measure relatedto contrast use weights related to the distance from the GLCMdiagonal, for example:

$$\text{Constract}=\sum_{i,j} p_{i,j} \times (i-j)2 \tag{1}$$

$$\text{Homogeneity}=\sum_{i,j} p_{i,j} /(1+(i-j)2) \tag{2}$$

$$\text{Dissimilarity}= \sum_{i,j} p_{i,j} \times |i-j| \tag{3}$$

And some features measure related orderliness. It means how regularthe pixel values with the window, for example:

$$\text{ASM}=\sum_{i,j} p_{i,j}2 \tag{4}$$

$$\text{Entropy}=\sum_{i,j} p_{i,j} \times \log(p_{i,j}) \tag{5}$$

$$\text{Max}=\max(p_{i,j}) \tag{6}$$

Where ASM is angular second moment. Energy is the square root ofASM. And MAX is the largest $p_{i,j}$ value found within thewindow.There are many other features too, for example entropy,correlation and difference variance etc.

The procedure for forming HOT features is shown inAlgorithm 1. Detection windows are samples. A set of sizeof N×N, e.g., 8×8 pixels, cells were obtained byslide windows and aligned in order to form a histogram$R=[R_1,R_2,\ldots R_M]$. GLCM is calculated at firstand the normalization is finished then. Afterwards the texturalfeatures are computed in every cell. The overlap cells are groupedover complete spatial region to form our final features.

Algorithm 1. Our Feature: GLCM-based Texture Histogram

1: Detection window is given and unit cell is obtained by slide window of size N×N pixels;
2: Calculate the $p_{i,j}(\theta,s)$ in four directions and let distance s=1 and normalization $p_{i,j}(\theta,s)$

$$pi,j(\theta,s)=\#\{((k,l)(m,n))\in (lx\times ly),I(k,l)=i,I(m,n)=j\}$$

3: Compute texture values by using $p_{i,j}$ to sign it as H(R) and form the HOT_cell:

$$HOT\_cell=[H1(R),H2(R),\ldots,Hk(R)]$$

Where K is the number of θ and every H(R) is a feature vector including six or more kinds of texture values.
4: Form a block from overlapped cells and normalize the every cell's values.
5: Output the final feature as from all the blocks.

Using above method, the shape outline information and detailedtexture information were ideally combined in a histogram. Thishistogram can describe the cat head feature effectively and can tellthem apart from all possible background images.

3 Experimental Analysis and Results

In human face detection, all training samples are normalized by arotation-scale transformation for reducing the variability andimproving the accuracy of a detector. In literature [16],authors show that the cat head cannot be normalized by a rotation +scale transform due to the large intra-class variation. Theynormalize all images by ears and by eyes separately. Every cat headimage is manually labeled by 9 points as shown in Figure 2. The author provides shape detector and texture detectorseparately too. Joint two detectors they obtained better performancethan use only one and they have good precision 0.91 for a fixedrecall 0.4.

In this paper we rotate and scale cat images by ears and eyes toobtain the opposite samples at first. The both ears and eyes are ona horizontal line and the distance between them are 36 pixels and 24pixels separately, and based on ears points, the head region arealigned and scaled to a base resolution of 48×48 pixels or 32×32 pixels based on eyes points. We will show thesimulation results bellow.

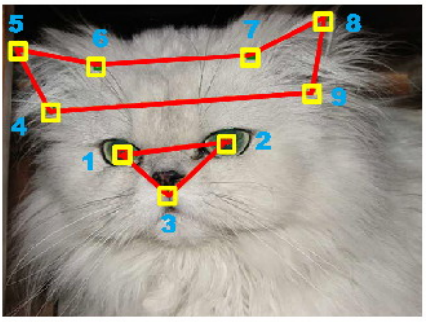


Fig. 2. The cat head image is manually labeled by 9 points

3.1 Data Set and Training Samples

Cat head samples were collected by extracting about 10,000 welllabeled cat images downloaded from http://mmlab.ie.cuhk.edu.hk/. Werotate and scale the images by the 9

points. Most of cat data arenear frontal view. The cat head samples do not include posevariation but a variety of face and illumination conditions. Werandomly divide the 10,000 samples into three sets: 4,000 fortraining and 2,000 for validation, and 4,000 for testing.

Negative samples were collected more than 10,000 images from the Webthat do not contain any of the objects of interest. Each image waspartitioned into patches of size48×48 pixels or 32×32 pixels and selected randomly 4 times so that the number of initialnegative samples became 4 times of the positive samples. There are40,000 negative samples for the cat head classifier and detector.

3.2 Performance Evaluation by Cross Validations

We adopt a 10-fold cross validation for performance evaluation andsvm-based classifier was used. The samples are partitioned into 10sets. Nine sets are used for training, and the trained classifier istested on the remaining set. The average performance is calculatedby repeating training and testing 10 times on different combinationsof samples sets. Error rates are calculated by counting the numberof misclassification.

3.3 Experiment: Based on Flicrk Cat Database

We test the performance of different textural features in different θand distance s. We choose the fourdirection: θ=[0°,45°,90°,145°], and distance s=[1,2,…L].Where Lis fixed on one fourth of the samplessize. The features vector is obtained by the texture values ofhomogeneity, dissimilar, contract, ASM, Energy, MAX and Entropy.

Figure 3 and figure 4 give thecomparison results of HOG and HOT by samples rotating by eyes andears separately.

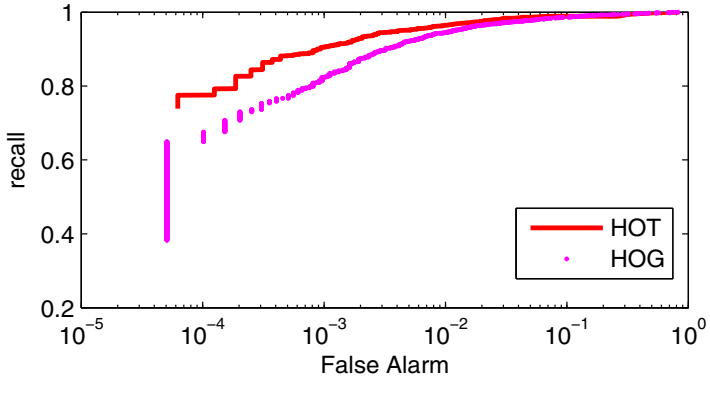


Fig. 3. Performance compared our feature with HOG

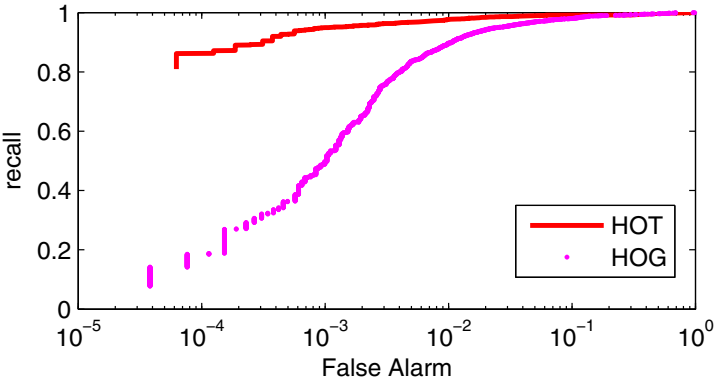


Fig. 4. Performance compared our feature with HOG

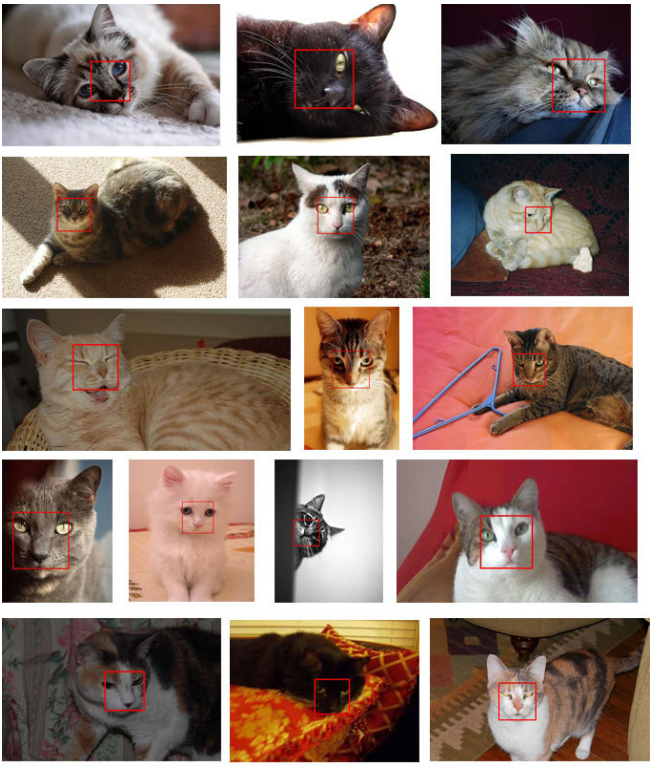


Fig. 5. Detection results

As the introduction shows, figure 1 (b) is the resultof cat head aligned by ears. It has more clear shape profile thantexture information. Figure 1(c) is the result of cathead

aligned by eyes. There is rich texture information in it. So,HOG has better performance in Figure 3 than inFigure 4. HOTdescriptor includes the cat head'stexture information and its histogram aligned method uses the shapeinformation, so it has more good performance than HOG in figure 3 and figure 4.

We also evaluate the proposed approach on the PASCAL 2007 cat data.Figure 5 gives some selection examples having variableappearance, head shape, illumination, and pos.

4 Conclusion

In this paper, we have proposed a novel feature, GLCM-based texturehistogram, for cat head detection. HOG is classical feature that hasgood performance on detecting human body by clearly and completelyexpressing the edge information of human body. But cat heads havebigger intra-variation and more rich textural information than humanbody. The features that we proposed can catch not only the edgeinformation but also textural information of images. So they havebetter performance than HOG. Experimental results show that theproposed features have more robust performance on illuminationchange and different scale size of eye or ears of cats.

Although we focus on the cat head detection problem in this paper, our approach can be extended to detect cat body or other categories of animals. In the future, we are planning to extend our approach to cat body detection and more animal categories. We are also interest in further improve the performance by using more effective detection mechanism.

Acknowledgments

The authors would like to thank Doctor Yan Shuicheng of SingaporeNational University for kindly providing all the test images. DoctorYan and other members of his Lab helped significantly in improvingthe ideas of this work.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients forhuman detection. In: CVPR, vol. 1, pp. 886–893 (2005)
2. Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T.: Fast human detection using a cascade of histograms of oriented gradients. In: CVPR, vol. 2, pp. 1491–1498 (2006)
3. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Machine Intell. 20(1), 23–38 (1998)
4. Viola, P., Jones, M.J.: Robust real-time face detection. Intl. Journal of Computer Vision 57(2), 137–154 (2004)
5. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR, vol. 1, pp. 878–885 (2005)
6. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. IEEE Trans. Pattern Anal. Machine Intell. 28, 1863–1868 (2006)

7. Papageorgiou, C., Poggio, T.: A trainable system forobject detection. Intl. J. Computer Vision 38, 15–33 (2000)
8. Schneiderman, H., Kanade, T.: A statistical method for 3dobject detection applied to faces and cars. In: CVPR, vol. 1, pp. 746–751 (2000)
9. Felzenszwalb, P.F.: Learning models for object recognition. In: CVPR, vol. 1, pp. 1056–1062 (2001)
10. Gavrila, D.M., Philomin, V.: Real-time object detection for smart vehicles. In: CVPR, vol. 1, pp. 87–93 (1999)
11. Lowe, D.G.: Object recognition from local scale-invariantfeatures. In: ICCV, vol. 2, pp. 1150–1157 (1999)
12. Levi, K., Weiss, Y.: Learning object detection from a smallnumber of examples: the importance of good features. In: CVPR, vol. 2, pp. 53–60 (2004)
13. Tuzel, O., Porikli, F., Meer, P.: Human detection viaclassification on Riemannian manifolds. In: CVPR (2007)
14. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learningshapelet features. In: CVPR (2007)
15. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet partdetectors. In: ICCV, vol. 1, pp. 90–97 (2005)
16. Zhang, W., Sun, J., Tang, X.: Cat Head Detection- How to effectively exploit Shape and Texture Features. In: Proceedings of ECCV-European Conference on Computer Vision, pp. 802–816 (2008)
17. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based objectdetection in images by components. PAMI 23, 349–361 (2001)
18. Schwartz, E.L.: Spatial mapping in the primate sensory projection:analytic structure and relevance to perception. Biological Cybernetics 25, 181–194 (1977)
19. Smith, J.R., Chang, S.F.: Automated binary texture feature setsfor image retrieval. In: Proc. of the IEEE International Conferenceon Acoustics, Speech, and Signal Processing, pp. 2239–2242 (1996)
20. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture Features for Image Classification. IEEE Trans. on Systems, Man and Cybernetics 3, 610–621 (1973)

Time Serial Model of Rock Burst Based on Evolutionary Neural Network

Wei Gao

Wuhan Polytechnic University
Wuhan 430023, Hubei, China
wgaowh@hotmail.com

Abstract. Rock burst is a mining disaster, which can bring large harm. The theory studies have showed that, rock burst is a kind of dynamic phenomenon of mining rock, and is a kind of dynamic disaster from mining. Considering the dynamic character of rock burst, the construction time serial model of rock burst is studied by evolutionary neural network based on immunized evolutionary programming. At last, the proposed model is tested by a real magnitude series of a mine. The result has showed that, evolutionary neural network model not only has high approaching precision, but also has high predicting precision, and is a good method to construct the non-linear model of rock burst.

Keywords: rock burst; mechanics of explosion; dynamic model; evolutionary neural network; time serial.

1 Introduction

In coal mining, at sometimes, the coal or rock will be burst, and a lot of energies will be released. At the same time, the nearby stratum will be produced violence shake, which will be earthquake produced be people. This phenomenon has been called rock burst or coal burst in coal mining branch. From the first rock burst happened in England at 1738, this disaster have been occurred in many coal mine in a lot of countries, such as USA, former U.S.S.R. and Poland, et al. In our country, from the first rock burst happened in Fushun mine at 1933, this disaster have become more and more frequently [1]. So, how to prevent this disaster has become an important work.

For its importance, there have been a lot of studies on rock burst. And many theories have been proposed [2], such as strength theory, stiffness theory, energy theory [3], burst trend theory, catas trophic theory [4], chaotic theory [5], fractal theory [6], et al. Also, a lot of methods to monitor rock burst have been found [7], such as drilled chip measurement, AE technique [8], seismological method [9], electromagnetic emission method [10], et al. We can find that, those theories have only studied the rock burst by mechanical method or mathematic method, and cannot used well in real engineering. While those monitor methods are all only to collect data about rock burst simply and not propose a method to analyze data from theory. So, a utility method to study rock burst is to analyze field data and extract essential model about rock burst.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 406–413, 2010.

2 Time Serial Model of Rock Burst Based on Neural Network

For rock burst system is a very complicated dynamic system, its time serial is a very complicated non-linear time serial. To model this kind of time serial, neural network is a very suitable one [11].

The general process of model construction by neural network is as follows.

(1) Supposing the time serial is $\{x(i),i=1,2, \ldots ,n\}$, before model construction, the sample set from the original serial that is applied in training and forecasting process must be constructed. As to forecasting problem, generally the network structure is multi-input and single-output model.

(2) After the above work, the training sets can be formed by sample modes from original serial. If neuron number of input layer is p, the number of the training sets will be n-p. The training sets are as follows,

The first sample is that input values are $x(1), \ldots ,x(p)$ and output value is $x(p+1)$; The second sample is that input values are $x(2), \ldots ,x(p+1)$ and output value is $x(p+2)$; … ; The sample which order is n-p is that input values are $x(n-p), \ldots ,x(n-1)$ and output value is $x(n)$.

In real practice, in order to get the better results, the rolling model method is used, which principles can be showed in Fig. 1. In this method, as a new serial value is produced, this value is added into time serial and the oldest one is displaced. So, the time serial that is used to construct neural network model is all new serial, and then the timeliness effect can be maintained.

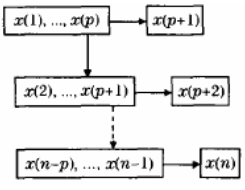


Fig. 1. Principles of rolling model for non-linear time serial forecasting

3 Time Serial Model of Rock Burst Based on Evolutionary Neural Network

To construct a neural network model for rock burst time serial, the construction of neural network is the main problem to be solved. Because for this problem, the hidden layer construction and input layer construction all can be confirmed. This problem can be solved by evolutionary algorithm very well. Here, as a primary study, one new evolutionary neural network which construction is confirmed by evolutionary algorithm and which weight is confirmed by MBP (Modified Back Propagation) algorithm is proposed. To make problem simpler and generalization bigger, the three layers neural network is studied. So, here, only the number of input neuron and number of hidden layer neuron are to be confirmed. In MBP algorithm, there are two parameters, iterating step η and inertia parameter α, to be confirmed. These two

parameters affected MPB algorithm very seriously. So, these two parameters are all confirmed by evolutionary algorithm. And then, in evolutionary neural network, there are four parameters to be evolved. In order to get the better effect, the new evolutionary algorithm- immunized evolutionary programming [12] proposed by author is used in evolutionary neural network. In order to study conveniently and directly, the indirect coding method is used in our algorithm. Also, to make the model as simple as possible, the full-linking network is applied, that is to say, the designed ENN must satisfy follow conditions.

(1) The input neuron has no input linking.
(2) The output neuron has no output linking.
(3) There is only one linking between any two neurons.
(4) There is only feed forward linking between any two neurons.

3.1 Immunized Evolutionary Programming

The main steps of immunized evolutionary programming are as follows.

(1) Individual Description: In immunized evolutionary programming, the individual is described directly by its representation. That is, if the search space is a *n*-dimension space, then the individual is described by *n*-dimension vector.

(2) Fitness Function: In immunized evolutionary programming, the fitness function of individual is generally a kind of transformation of its objective function. The purpose of this transformation is not only to guarantee the fitness value of each individual positive, but also to maintain the diversity among individuals.

(3) Mutation Operator: Mutation operation is the only optimization search operator in immunized evolutionary programming.

In traditional evolutionary programming, the mutation operator is one kind of Gauss mutation. While in immunized evolutionary programming, one kind of adaptive Cauchi mutation is used, which form is as follows,

Suppose one individual of the population is $X = (a_1,...,a_4)$, the new mutated individual is $X' = (a_1',...,a_4')$, then the component of the new individual can be described as follows,

$$a_i' = a_i + \sigma_i \cdot T \cdot C_i(0,1) \qquad (i=1\text{-}4). \tag{1}$$

where,

$$\sigma_i = 1.0 \Big/ \sqrt{\beta_i F(X) + \gamma_i} \qquad (i=1\text{-}4). \tag{2}$$

where, σ_i is the standard deviation of the parameters, $C_i(0,1)$ is the Cauchi random number; $F(X)$ is the fitness of individual, β_i, γ_i are two parameters, there $\beta_i = 1$, $\gamma_i = 0$, T is a adaptive parameter, which description is as follows,

$$T = \frac{T_0}{F_{\max} - F_{avr}} \cdot \tag{3}$$

where, T_0 is the initial parameter, which value is 2.5; $F_{\max}$, $F_{\min}$ are the maximum and minimum fitness of the current population.

The adaptive mutation can make the disturbing extent of mutation adaptively changeable through the evolutionary extent, so the search performance of the whole algorithm can be improved.

(4) Selection Operator: In traditional evolutionary programming, the selection operator is a kind of stochastic tournament modal that is called random competition with parameter q. The basic process of this operator is as follows.

a. Calculate the fitness of each offspring.

b. Conduct pairwise comparison over the union of parents and offspring. For each individual, q opponents are chosen uniformly at random from all the parents and offspring. For each comparison, if the individual's fitness is no smaller than the opponent's, it receives a "win."

c. Select the individuals out of the parents and offspring that have the most wins to be parents of the next generation.

Considering the shortcomings of losing diversity of population and premature of algorithm, in immunized evolutionary programming, we improve the selection operation of evolutionary programming by principles of artificial immune system. From principles of artificial immune system, we take the problem to be solved as antigen and solution as antibody. And then, the adjustment mechanism based on thickness is introduced to improve selection operation.

According to the stimulative reaction and restraining reaction in immune system, one kind of thickness factor is introduced into selection operation to adjust the score of individual. If the thickness of individual is high, the premature is easy to appear in the evolutionary process, so the individual whose thickness is high should be restrained and the selection probability of individual whose fitness is large must be high. The detailed method is as follows, the adjustive modification based on thickness and fitness of individual is added to the score of individual, which is as follows.

$$p'(i).scores = p(i).scores + \alpha \cdot C \cdot (1 - \frac{F(i)}{\max F}) \cdot p(i).score + \beta \cdot \frac{F(i)}{\max F} \cdot p(i).scores \,. \tag{4}$$

where, α, β are parameters which are in the range of [0~1], and here $\alpha = \beta = 0.5$, C expresses the thickness of individual, whose definition is the ratio of number of individuals whose fitness is highest or near highest to total number of population, and here it can be expressed as follows,

$$C = \frac{t \cdot (0.8 \cdot \max F \to \max F)}{N} \,. \tag{5}$$

where, numerator is the summation of individuals whose fitness is between max$F$ and 0.8max$F$, $F(i)$ is the fitness value of individual whose sequence number is i, max$F$ if the maximum fitness of the population.

From the score formula equation, we can see that, as to the individuals whose thickness is high and total number is t, if their fitness is high, their scores are little (this can be seen from second term of score formula). While if the thickness of

individual is low, the score of individuals whose fitness is high must be higher due to the third term of score formula.

3.2 New Evolutionary Neural Network

The details of new evolutionary neural network are as follows.

(1) The search range of input neuron and hidden layer neuron are given firstly. And also the search range of two parameters in MBP algorithm are given. And some evolutionary parameters, such as evolutionary generation stop criteria, individual number in one population, the error criteria of evolutionary algorithm, number of output neuron in neural network, iterating stop criteria and iterating error criteria in MBP algorithm are all given.

It must be pointed out that, to construct the suitable samples, the number of input neuron must be smaller than total number of time series.

(2) One network construction is generated by two random numbers in search range of input neuron and hidden layer neuron. And also, one kind of MBP algorithm is created by two random numbers in search range of parameters η and α. And then, one individual can be generated by the four parameters.

It must be pointed out that, for two numbers of neuron are integer numbers and two MBP parameters are real numbers, so the expressions of one individual must be structural data.

(3) To one individual, its fitness value can be gotten by follow steps.

a. The whole time series of stock market is divided to construct the training samples based on number of input neuron and number of hidden layer neuron. And also, the total number of samples is noted.

b. The whole learning samples are to be divided into two parts. One part is the training samples, which is to get the non-linear mapping network. The other part is the testing samples, which is to test the generalization of network.

c. The initial linking weights of network individual are generated.

d. The iterating step of MBP algorithm is taken as $j = 1$.

e. This network individual is trained by testing samples, and the square error $E(j)$ is computed, and this error is taken as minimum error of the whole training, $\min E = E(j)$. If min$E$ is smaller than the error criteria of evolutionary algorithm, then the fitness value is min$E$. And the computing process is changed to step (3).

f. This network individual is trained by training samples. If its training error is smaller than iterating error criteria of MBP algorithm, then the fitness value is also the min$E$. And the computing process is changed to step (3).

g. The whole linking weights are adjusted by MBP algorithm.

h. $j = j + 1$, and the computing process is changed to step e.

i. If j is lager than iterating stop criteria of MBP algorithm, then the fitness value is also the min$E$. And the computing process is changed to step (3).

(4) If the evolutionary generation reaches its stop criteria or computing error reaches error criteria of evolutionary algorithm, then the algorithm stop. At this time, the best individual in last generation is the searching result.

(5) Every individuals in population are mutated. For there are different data types in one individual, the different mutation types are used for each parameter. For

numbers of input neuron and hidden layer neuron are integer number, the uniform mutation is used. For parameters η and α are real numbers, the adaptive Cauchi mutation is used. And then the offspring population is generated.

(6) The fitness value of each individual in offspring population is calculated by the method in step (3).

(7) The set of offspring population and parent population is selected by selection operation based on thickness, then the new offspring population is generated.

(8) The number of evolutionary generation increases 1, then the computing process is changed to step (4).

From the above algorithm, the four parameters, number of input neuron, number of hidden layer neuron, two parameters η and α in MBP algorithm can be confirmed. So, the optimization neural network for rock burst time serial will be gotten.

4 Engineering Example

To verify the method proposed in this paper, the rock burst data from Daanshan mine of Beijing mine bureau in reference [13] is used here. The rock burst has been studied by seismological method in Beijing mine bureau. But for the rock burst is a microseismic activity, there will be exist large error to study rock burst by seismological method. So, to solve this problem, here the rock burst is studied by our method.

The microseismic magnitude data in Daanshan mine of Beijing mine bureau are used here, which are showed in follow Table 1.

Table 1. The comparison of computing results and magnitude records

Serial number	1	2	3	4	5	6	7	8	9
Observed value	1.0	2.2	1.4	1.1	1.6	1.3	1.2	1.0	2.1
Method in [13]	1.2	2.2	1.37	1.07	1.7	1.28	1.22	1.1	2.17
				Fitting results					
Our method	1.08	2.14	1.41	1.18	1.52	1.34	1.15	0.91	2.02

Serial number	10	11	12	13	14	15	16	17	18
Observed value	1.5	1.5	2.1	1.1	2.1	1.6	1.1	1.9	2.6
Method in [13]	1.62	1.44	2.3	1.11	2.3	1.66	1.01	1.75	2.74
		Fitting results				Forecasting results			
Our method	1.61	1.53	2.21	1.15	2.17	1.56	1.12	1.95	2.7

The former 13 data are used for training computation, and the last 5 are used for testing computation. And the evolutionary neural network model is trained by the training data and testing data. In computation, the parameters are set as follows. The range of input neuron number is from 1 to 10. The range of hidden neuron number is from 1 to 20. The range of iterating step is from 0.01 to 1. The range of inertia parameter is from 0 to 0.99.

From the computation of evolutionary neural network, the follow results can be gotten. The number of input neuron is 4. The number of hidden neuron is 7. The iterating step is 0.226 and inertia parameter is 0.948. So, the optimal construction of neural network for rock burst serial model is 4-7-1, that is to say, the training samples can be constructed as follow Fig. 2.

Input data				Output data
(1.0	2.2	1.4	1.1	1.6)
(2.2	1.4	1.1	1.6	1.3)
(1.4	1.1	1.6	1.3	1.2)
(1.1	1.6	1.3	1.2	1.0)
(1.6	1.3	1.2	1.0	2.1)
(1.3	1.2	1.0	2.1	1.5)
(1.2	1.0	2.1	1.5	1.5)
(1.0	2.1	1.5	1.5	2.1)
(2.1	1.5	1.5	2.1	1.1)

Fig. 2. Training samples

The first seven sets of samples are used as training samples, and the last two as testing samples.

The whole data are computed by the above neural network, the follows results can be gotten in Table 1. From Table 1, we can see that, evolutionary neural network model not only has high approaching precision, but also has high predicting precision, and is a good method to construct the non-linear model of rock burst.

5 Conclusions

The theory and practice are all proved that, to study complicated system, using its external behavioral representation is a very good method. For the neural network is a black box model, which can model very complicated system very well, in this paper, one method based on new evolutionary neural network to study rock burst time serial is proposed. From used in one real mine engineering, our method is tested to be a very good method, which can not only fit measured data well, but also can forecast accurately.

For rock burst system is a very complicated dynamic system, which cannot be described very well by one system parameter. In this paper, limited by technical condition, the proposed method is not very comprehensively. So, to study rock burst system by more system parameters is our next work.

Acknowledgments. The financial support from The National Natural Science Foundation of China under Grant No. 40872187 is gratefully acknowledged.

References

1. Zhao, B.J., Teng, X.J.: Rock Burst and Its Prevention. Mining Industry Press, Beijing (1995)
2. Zhou, X.J., Xian, X.F.: Research advance on rockburst theory and its engineering application in collieries. Journal of Chongqing University 21, 126–132 (1998)
3. Zhang, M.T.: Numerical simulation and theory study of rockburst. Chinese Journal of Rock Mechanics and Rock Engineering 6, 197–204 (1987)
4. Pan, Y.S., Zhang, M.T.: The study of coalburst by catas trophic theory. Journal of Fuxin Mining Institute 11, 12–18 (1992)
5. Song, W.Y., Pan, Y.S., Su, R.H., et al.: Forecast and chaos model of rockburst. Journal of China Coal Society 26, 26–30 (2001)
6. Li, Y., Huang, M., Zhang, L.C., et al.: Fractal dimension in controlling of coal outburst. Rock and Soil Mechanics 15, 34–38 (1994)
7. Tian, M.G., Liu, T.C.: Mining Pressure and Its Control. Mining Industry Press, Beijing (1992)
8. Qi, Q.X., Li, S.B., Wang, S.K.: Application of AE technique in monitoring ground pressure. Journal of China Coal Society 19, 221–232 (1994)
9. Zhang, S.Q., Zhang, Z.P., Yang, M.Y., et al.: The seismological research and development on mining tremors. Earthquake Research in China 9, 1–8 (1993)
10. Nie, B.S., He, X.Q., Wang, E.Y., et al.: Forecasting Rockburst with the Electromagnetic Emission Method in Coal Mine. Journal of Taiyuan University of Technology 31, 609–611 (2000)
11. Gao, W.: Study on forecasting method of nonlinear signal processing. In: Proceedings of 7th International Conference on Electronic Measurement and Instruments, Beijing, pp. 593–597 (2005)
12. Gao, W., Zheng, Y.R.: An new evolutionary back analysis algorithm in geotechnical engineering. Chinese Journal of Rock Mechanics and Rock Engineering 22, 192–196 (2003)
13. Liu, J.J., Wang, J.S., Liang, B., et al.: The gray system method in the magnitude forecast of rockburst. Journal of Liaoning Technical University 17, 222–224 (1998)

Multilayer Perceptron Network with Modified Sigmoid Activation Functions

Tobias Ebert, Oliver Bänfer, and Oliver Nelles

University of Siegen,
Department of Mechanical Engineering,
D-57068 Siegen, Germany
tobias.ebert@uni-siegen.de

Abstract. Models in today's microcontrollers, e.g. engine control units, are realized with a multitude of characteristic curves and look-up tables. The increasing complexity of these models causes an exponential growth of the required calibration memory. Hence, neural networks, e.g. multilayer perceptron networks (MLP), which provide a solution for this problem, become more important for modeling. Usually sigmoid functions are used as membership functions. The calculation of the therefore necessary exponential function is very demanding on low performance microcontrollers. Thus in this paper a modified activation function for the efficient implementation of MLP networks is proposed. Their advantages compared to standard look-up tables are illustrated by the application of an intake manifold model of a combustion engine.

Keywords: nonlinear system identification, neural network, multilayer perceptron network, modified activation function.

1 Introduction

The requirements of modern vehicles with respect to fuel consumption, emissions and driveability can be fulfilled only with increasing variability of the combustion engine. The corresponding actuators are driven by the engine control unit (ECU), which requires models of growing complexity. For the model of the cylinder air mass an exemplary comparison of manifold absolute pressure (map) based methods and a neural network is performed. Both approaches benefit from taking into account prior knowledge about the physics of the modelled system for the model generation. While map-based methods are the best choice for simple systems with few degrees of freedom, the modeling effort grows exponentially with increasing complexity for this method. Thus, for complex models, i.e. for models with more than five inputs, neural networks such as multilayer perceptron (MLP) networks are superior. Furthermore, some functions in the ECU, e.g. the intake manifold model, face the challenge that the outputs of the model have to be computed in real-time, because the results are necessary for each cycle of the combustion engine. Thus an efficient implementation of the MLP on the inbuilt microcontroller is of particular importance. The computation of

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 414–421, 2010.

the standard activation function (AF) is not possible, because the calculation of an exponential function is too demanding for the microcontroller. The only way to implement this standard AF could be realized by storing this function in a look-up table (LUT). A more promising method is to replace the sigmoid with a polynom of similar form. This has the crucial advantage that the complex computation of the exponential function is reduced to a simple multiplication for the polynomial function.

This article is organized as follows. Section 2 gives an overview on multilayer perceptron networks and specifies their advantages. In Sect. 3 the standard sigmoid activation function and a modified activation function, consisting of piecewise polynomial functions, are introduced and their properties are discussed. The modified activation function is implemented in an intake manifold model and the results are discussed in Sect. 4. This paper ends by summarizing the important conclusions.

2 Multilayer Perceptron Networks

A MLP network is an universal approximator [5] and consists of simple calculation units, the neurons, and directed, weighted connections between them. These neurons are arranged in so called layers. Usually a perceptron network consists of a layer inputs and one or more neuron layers with trainable weights. A perceptron network with a single trainable layer can divide the input space by means of a hyperplane. A two-stage perceptron can classify convex polygons. Finally a three-stage perceptron can classify sets of any form by combining and separating arbitrarily many convex polygons. Hence a MLP with one input, two hidden and one output layer, as shown in Fig. 1, further described in [6], is a common form.

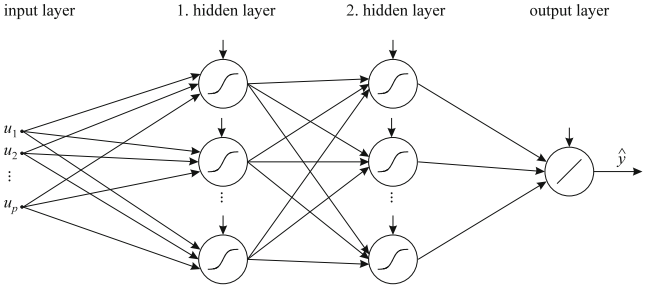


Fig. 1. A feedforward MLP Network with three trainable layers: Two hidden layers and one output layer

Figure 2 shows a neuron of the hidden layer of a multilayer perceptron. This single neuron is called a *perceptron*. The operation of this neuron can be splited into two parts. First, ridge construction is used to project the inputs $\boldsymbol{u} = [u_1 \ u_2 \ \cdots \ u_p]^{\mathrm{T}}$ on the weights. In the second part, the nonlinear activation function $g(x)$ transforms the projection result. The perceptron depicted

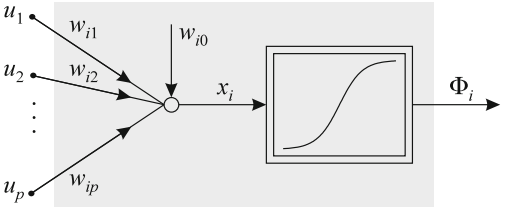


Fig. 2. A perceptron: The ith hidden neuron of a multilayer perceptron

in Fig. 2 depends on nonlinear hidden layer parameters. These parameters are called *hidden layer weights*:

$$\boldsymbol{\theta}_i^{(\mathrm{nl})} = [w_{i0}\ w_{i1}\ w_{i2}\ \cdots\ w_{ip}]^{\mathrm{T}}\,. \tag{1}$$

The weights w_{i0} realize an offset, sometimes called "bias" or "threshold". These hidden layer weights determine the shape of the basis functions.

With M_1 and M_2 as the number of neurons in the first and second hidden layer, respectively, w_i as output layer weights, and $w_{jl}^{(1)}$ and $w_{ij}^{(2)}$ as weights of the first and second hidden layer, the basis function formulation becomes

$$\hat{y} = \sum_{i=0}^{M_2} w_i \Phi_i \left(\sum_{j=0}^{M_1} w_{ij}^{(2)} \xi_j \right) \qquad \text{with } \Phi_0(\cdot) = 1\,, \tag{2}$$

and with the outputs of the first hidden layer neurons

$$\xi_j = \Psi_j \left(\sum_{l=0}^{p} w_{jl}^{(1)} u_l \right) \quad \text{with } \Psi_0(\cdot) = 1 \text{ and } u_0 = 1\,. \tag{3}$$

Usually the activation functions of both hidden layer Φ_i and Ψ_j are chosen to be off saturation type, see Sect 3. With p as the number of inputs and q as the number of outputs, the number of trainable weights of a MLP with two hidden layers is

$$M_1\,(p+1) + M_2\,(M_1+1) + (M_2+1)\,q\,. \tag{4}$$

Owing to the term $M_1 M_2$ the number of weights grows quadratically with an increasing number of hidden layer neurons.

3 Activation Function

Typically, the activation function of hidden layer neurons is chosen to be of saturation type. Common choices are sigmoid functions such as the hyperbolic tangent

$$g(x) = \tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}\,. \tag{5}$$

This function exhibts the interesting property that the derivative can be expressed as a simple function of its output:

$$\frac{d\,\Phi_i}{dx} = \frac{d\,\tanh(x)}{dx} = \frac{1}{\cosh^2(x)} = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)}$$
$$= 1 - \tanh^2(x) = 1 - \Phi_i^2\,. \tag{6}$$

These derivatives are required in any gradient-based optimization technique applied for training of a MLP network; see Sect 4.1. The gradient with respect to the weights of the second hidden layer and $\tanh(\cdot)$ activation function is ($i = 1, \ldots, M_2$, $j = 0, \ldots, M_1$):

$$\frac{\partial \hat{y}}{\partial w_{ij}^{(2)}} = w_i(1 - \Phi_i^2)\Psi_j \qquad \text{with } \Psi_0 = 1\,. \tag{7}$$

The gradient with respect to the weights of the first hidden layer and $\tanh(\cdot)$ activation function is ($j = 1, \ldots, M_1$, $l = 0, \ldots, p$)

$$\frac{\partial \hat{y}}{\partial w_{jl}^{(1)}} = \sum_{i=0}^{M_2} w_i(1 - \Phi_i^2)w_{ij}^{(2)}(1 - \Psi_j^2)u_l \quad \text{with } u_0 = 1\,. \tag{8}$$

The gradient with respect to the ith *output layer weight* of an MLP is ($i = 0, \ldots, M$):

$$\frac{\partial \hat{y}}{\partial w_i} = \Phi_i \qquad \text{with } \Phi_0 = 1\,. \tag{9}$$

Note that the basis functions Φ_i in the above equations depend on the network inputs and the hidden layer weights. These arguments are omitted for better readability.

3.1 Modified Activation Function

The problem in the transition from a standard MLP network to a MLP on a microcontroller (μC) is to overcome the limitations μC impose on MLP calculations. It is important to distinguish two different problems: The minor problem is the limited computational performance due to the architecture of the μC. Thus calculations may have to be done with a reduced accuracy e.g. 16-bit. Early results show, that this limitation has an observable impairing effect on the resulting model, but it does not render the method useless.

The focus of this work and the major problem however is how to compute the network output in real time on a μC with low performance. To calculate a MLP output a series of calculations have to be done. While basic calculations e.g. addition, subtraction, multiplication and division are uncomplicated, there is one bottleneck, which cannot be solved in real time on a low performance μC: The calculation of the exponential function in the hyperbolic tangent. To overcome this, a suitable approximation has to be found. Traditionally a look-up table (LUT) with linear interpolation would be used in an ECU to do this.

An accompanying flaw of this is that a high-resolution LUT would be needed to achieve a satisfying accuracy and smooth model-transition, which demands a corresponding high consumption of data memory. Therefore as a trade-off between computing time and accuracy an approximation with piecewise polynomial models (PPM) is explored.

A simple approximation of the sigmoid function is possible by the application of two quadratic functions, given by

$$\hat{g}(x) = \begin{cases} -1 & \text{for: } x \leq -3, \\ \frac{(3\,m-1)}{9}\,x^2 + m\,x & \text{for: } -3 < x \leq 0, \\ \frac{(1-3\,m)}{9}\,x^2 + m\,x & \text{for: } 0 < x \leq 3 \text{ and} \\ 1 & \text{for: } x \geq 3. \end{cases} \tag{10}$$

The aim is to minimize the quadratic difference between the hyperbolic tangent and the approximation above. Numeric optimization gives us $m \approx 0.875$. See Figure 6a.

4 Application Intake Manifold Model

As a complementary work to [2] the realization of a multilayer perceptron network in a resource-friendly algorithm has been investigated for the application of the intake manifold model, see Fig. 3. For the abidance by the ambitious emission laws, the ECU has to model an exact air mass flow depending on the engine operating point. The needed injection fuel mass is calculated with the aid of the model in such a way that an optimal air fuel ratio ($\lambda = 1$) is reached for the exhaust aftertreatment in the catalyzer. Today the modeling of the air mass flow (MAF) depending on the manifold air pressure (MAP) occurs in a linearized form with the ordinate value OFFSET and the gradient SLOPE [4]. These two parameters are stored for steady state in the ECU with a set of numerous look-up tables subject to the engine speed N, the intake camshaft position CAM_IN, the exhaust camshaft position CAM_EX and the two-stage actuator settings for the swirl flap PORT and for the variable intake manifold VIM, see Fig. 4 [1].

4.1 Multilayer Perceptron Network Training

On the engine test bench, 360 operating points are measured by systematic combination of all inputs to calibrate the look-up table:

- 10 values for engine speed (N)
- 3 values for intake camshaft position (CAM_IN)
- 3 values for exhaust camshaft position (CAM_EX)
- 2 values for swirl flap position (PORT)
- 2 values for variable intake manifold length (VIM)

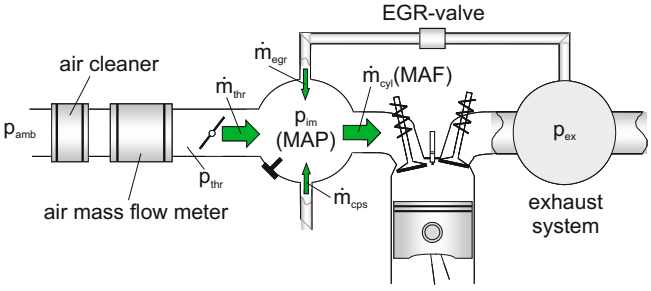


Fig. 3. Intake Manifold Model

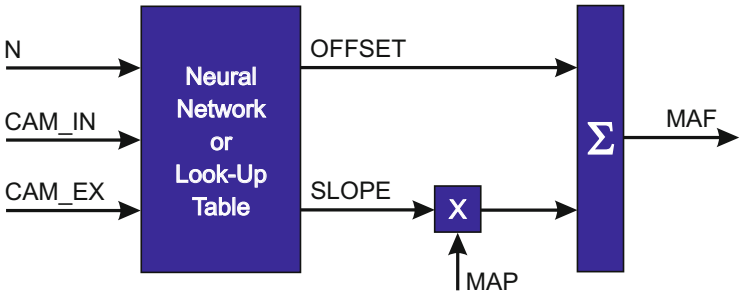


Fig. 4. Signal flow for one PORT-VIM-combination

Figure 5 shows the measured set values for one PORT-VIM-combination. The two-stage actuator settings PORT and VIM are binary values and therefore they are either zero or one. Hence, it proves successfully not to define both binary set values as an input for the neural network, but to train one network for each PORT-VIM-combination. The three inputs (N, CAM_IN and CAM_EX) and the two outputs (OFFSET and SLOPE) build the training data. With the aid of these data, a look-up table can be generated or a MLP network can be trained. After initialization of the weights recommended by Nguyen and Widrow [7] the MLP network was trained using the Levenberg-Marquardt algorithm [3].

The MLP network is extremely sensitive to its activation function. Hence there will be a problem if the approximation of the AF allows overshooting outside of the bounds of the standard hyperbolic tangent [-1,1]. To sort out this problem a more complex approximation has to be used, which strictly forbids overshooting. A solution to this is to increase the number and order of the polynomial functions. A sufficient increase consists of three polynomials of third order; see Fig. 6b.

4.2 Results

In Fig. 7a the overfitting problem is clearly distinguishable, since the measured data samples for the engine speed (1000 rpm, 2000 rpm, 3000 rpm, 4000 rpm,

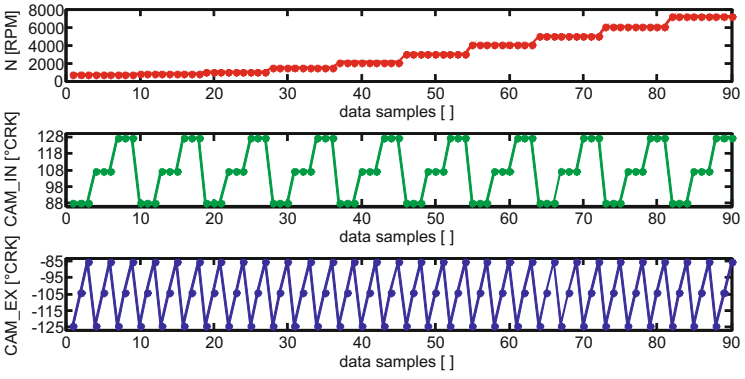


Fig. 5. Training input data for one PORT-VIM-combination

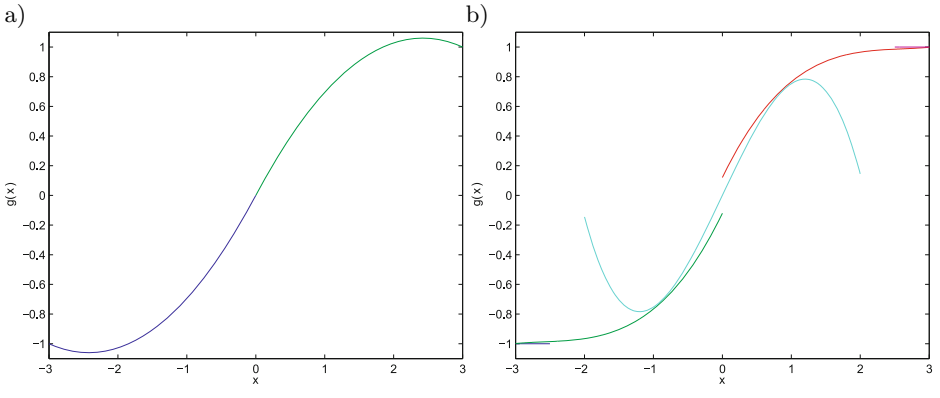


Fig. 6. a) Modified sigmoid function consisting of two polynoms of second order. b) Approximation of the activation function without overshooting.

5000 rpm, 6000 rpm) are approximated very well. The interpolation behaviour shows no satisfying results. Due to the fact that no validation data are obtainable, the analysis of the neural network interpolation behaviour is done by the comparison of interpolated values of MLP network and look-up table.

If the MLP is allowed to use the bounded approximation, the interpolation behaviour of the MLP in Fig. 7b will be almost equal to that of the LUT. Thus the bounded approximation is helpful to improve the interpolation behaviour. As a drawback in comparison to [2] a MLP network needs many more training attempts until a reliable solution is found. This leads to an extreme time consuming network training compared to a local model tree technique.

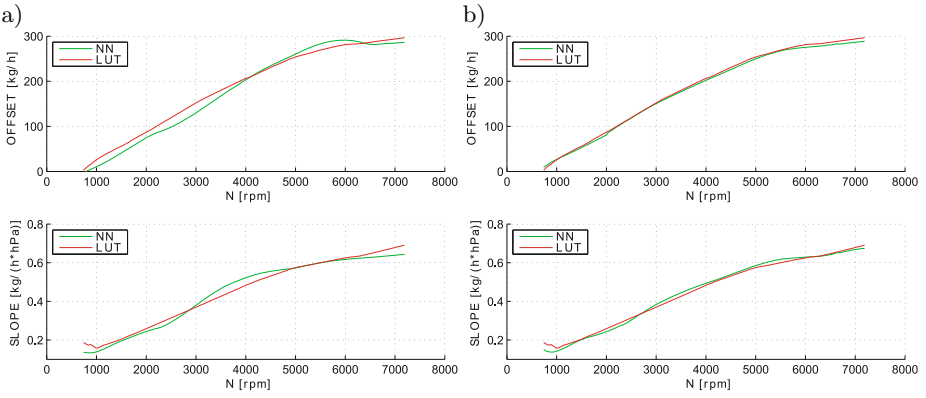


Fig. 7. a) Interpolation behavior with CAM_IN = 90°CRK, CAM_EX = −90°CRK and unbounded approximation of the sigmoid function. b) Interpolation behavior with CAM_IN = 90°CRK, CAM_EX = −90°CRK and bounded approximation of the sigmoid function.

5 Conclusion

With the aid of modified activation functions, it is achieved to replace all look-up tables of the intake manifold model with four different MLP networks, one for each combination of the two-stage actuator settings for the port flap and the variable intake manifold. The interpolation behaviour of the MLP network shows no indication of overfitting in comparison with the LUT interpolation. The realization in a series-ECU yields an obvious reduction of the necessary read-only memory at the expense of a moderately higher computational effort. The number of required parameters of a MLP does not grow exponentially, but as a good approximation linearly with the number of inputs. Hence, this alternative way of modeling avoids the so-called "curse of dimensionality".

References

1. Bänfer, O., Nelles, O., Kainz, J., Beer, J.: Local model networks - the prospective method for modeling in electronic control units? AT Zelektronik 3(6), 36–39 (2008)
2. Bänfer, O., Nelles, O., Kainz, J., Beer, J.: Local Model Networks with Modified Parabolic Membership Functions. In: 2009 International Conference on Artificial Intelligence and Computational Intelligence, pp. 179–183. IEEE, Los Alamitos (2009)
3. Hagan, M.T., Menhaj, M.B.: Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5(6), 989–993 (1994)
4. Heywood, J.B.: Internal Combustion Engine Fundamentals. McGraw-Hill, Inc., New York (1988)
5. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359–366 (1989)
6. Nelles, O.: Nonlinear System Identification. Springer, Berlin (2001)
7. Nguyen, D., Widrow, B.: Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: Proceedings of the International Joint Conference on Neural Networks, Washington, vol. 3, pp. 21–26 (1990)

Kernel Oblique Subspace Projection Approach for Target Detection in Hyperspectral Imagery

Liaoying Zhao[1], Yinhe Shen[1], and Xiaorun Li[2]

[1] Institute of Computer Application Technology, HangZhou Dianzi University, Hangzhou 310018, China
[2] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China
Zhaoly@hdu.edu.cn

Abstract. In this paper, a kernel-based nonlinear version of the oblique subspace projection (OBSP) operator is defined in terms of kernel functions. Input data are implicitly mapped into a high-dimensional kernel feature space by a nonlinear mapping, which is associated with a kernel function. The OBSP expression is then derived in the feature space, which is kernelized in terms of the kernel functions in order to avoid explicit computation in the high-dimensional feature space. The resulting kernelized OBSP algorithm is equivalent to a nonlinear OBSP in the original input space. Experimental results based on simulated hyperspectral data and real hyperspectral imagery shows that the kernel oblique subspace projection (KOBSP) outperforms the conventional OBSP.

Keywords: kernel function; kernel oblique subspace projection; hyperspectral image.

1 Introduction

Hyperspectral imaging (HSI) sensors and their applications have been developed tremendously over the past three decades. Detection of man-made objects or materials (i.e. targets) by using hyperspectral sensors is of great interest for civilian and military applications. Linear spectral mixture models are widely used in hyperspectral image for detection and identification of a desired target spectral signature in spectrally mixed pixels of subpixels. Several techniques based on linear spectral mixture model have been proposed in the literature, such as the Reed-Xiaoli anomaly detector (RX) [1], the orthogonal subspace projection (OSP) [2], matched subspace detector (MSD) [3], spectral matched filter (SMF) [4], and adaptive subspace detectors(ASD) also known as adaptive cosine estimator (ACE) [5,6]. For the OSP to classify the pixel or to detect the target effectively, a complete knowledge of the signature abundances is required. Generally, it is difficult to obtain such prior knowledge in many practical applications. Chein-I Chang extends the OSP by three unconstrained least squares subspace projection approaches, called signature space OSP (SSP), target signature space OSP (TSP), and oblique subspace projection (OBSP)[7].

In real imagery, when light reflects off surfaces that are composed of an intimate mixture of various material components that cause multiple bounces, spectral mixing

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 422–431, 2010.

tends to become nonlinear.[8] The nonlinear relationships between different spectral bands within the target or clutter spectral signature need to be exploited in order to better distinguish between target and background. Though the above methods have been widely used for hyperspectral image classification and target detection, they only exploit second order correlations, thus completely ignoring nonlinear (higher order) spectral interband correlations that could be crucial to discriminate between target and background. To overcome this drawback, Heesung Kwon and Nasser M. Nasrabadi introduced nonlinear versions of RX, MSD, OSP, SMF and ASD detectors which effectively exploit the higher order spectral interband correlations in a high-(possibly infinite) dimensional feature space associated with a certain nonlinear mapping via kernel-based learning methods [9,10]. Based on the ideas of SSP and kernel-based learning theory, we proposed a kernel version of the SSP called as kernel signature orthogonal space projection (KSSP) [11]. A nonlinear mapping of the input data into a high dimensional feature space is often expected to increase the data seperability and reduce the complexity of the corresponding data structure [12]. Certainly, the use of kernel methods offers many advantages: they combat the high dimensionality problem in hyperspectral images, make the method robust to noise, and allow for flexible non-linear mappings with controlled (regularized) complexity.

Motivated by the research on the above kernel-based techniques, the aim of this paper is to extend OBSP to its corresponding kernel versions by using the idea of kernel-based learning theory. The rest of the paper is outlined as follows. Section 2 provides the background to the kernel-based learning methods and kernel trick. Section 3 introduces the linear OBSP and its kernel version. Performance comparison between the conventional and the kernel version of the algorithm is provided in Section 4. Conclusions are given in Section 5.

2 Kernel Methods and Kernel Trick

Suppose that the input hyperspectral data is represented by the data space χ and $\mathcal{F}$ is a feature space associated with χ by a nonlinear mapping function ϕ

$$\phi : \chi \to \mathcal{F}, \boldsymbol{x} \to \phi(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ is an input vector in χ which is mapped into a potentially much higher (could be infinite) dimensional feature space.

Definition 1. (Kernel function) A kernel is a function k that for all $\boldsymbol{x}, z \in \chi$ satisfies

$$k(\boldsymbol{x}, z) = < \phi(\boldsymbol{x}) \cdot \phi(z) > \tag{2}$$

where $< \cdot >$ is the dot products.

Due to the high dimensionality of the feature space $\mathcal{F}$, it is computationally not feasible to implement any algorithm directly in feature space. However, kernel-based learning algorithms use an effective kernel trick given by (2) to implement dot products in feature space by employing kernel functions [12]. There are a large number of Mercer kernels, such as the Gaussian radial basis function (RBF) kernel, polynomial

kernel, sigmoidal kernel,etc., that have the kernel trick property (see [12] for detailed information about the properties of different kernels and kernel-based learning).

3 OBSP and Kernel OBSP Algorithm

3.1 Linear Mixture Model and OBSP

Linear spectral unmixing is a widely used approach in multispectral/hyperspectral imagery to determine and quantify individual spectral signatures in a mixed pixel. Let $\boldsymbol{r}$ be a mixed pixel consisting of L spectral bands and p spectral signature present in the mixed pixel. $\boldsymbol{r}$ is denoted by an $L\times 1$ column vector, and the p endmembers are represented by an $L\times p$ matrix $\boldsymbol{M}$ whose columns are the endmember spectrums. A linear mixture model for pixel $\boldsymbol{r}$ is described by

$$\boldsymbol{r}=\boldsymbol{M}\boldsymbol{\alpha}+\boldsymbol{n} \tag{3}$$

where α is a $p\times 1$ column vector whose elements are the coefficients that account for the proportions (abundances) of each endmember spectrum contributing to the mixed pixel and n is an $L\times 1$ vector representing an additive zero-mean Gaussian noise with covariance matrix $\sigma^2\mathrm{I}$ (σ^2: noise variance) and I is an $L\times L$ identify matrix.

In the following, we briefly review the OBSP approach given in [7]. First of all, we rewrite model (3) as

$$\boldsymbol{r}=\boldsymbol{d}\alpha_i+\boldsymbol{U}\boldsymbol{\gamma}+\boldsymbol{n} \tag{4}$$

where $\boldsymbol{M}$ is made up of a desired signature $\boldsymbol{d}=\boldsymbol{m}_i, 1\le i\le p$ and the undesired spectral signature matrix $\boldsymbol{U}=(\boldsymbol{m}_1\cdots\boldsymbol{m}_{i-1}\,\boldsymbol{m}_{i+1}\cdots\boldsymbol{m}_p)$, while α_i is the abundance of $\boldsymbol{m}_i$ and $\boldsymbol{\gamma}=(\alpha_1\cdots\alpha_{i-1}\,\alpha_{i+1}\cdots\alpha_{p-1})^T$ represent the abundances for the undesired spectral signatures. Model (4) can be extended straightforwardly to more than one desired signature. We also assume that $<\boldsymbol{d}>$, $<\boldsymbol{M}>$ and $<\boldsymbol{U}>$ are the spaces linearly spanned by $\boldsymbol{d}$, $\boldsymbol{M}$ and $\boldsymbol{U}$, respectively.

In the SSP and OBSP, model (4) was reformulated by a posteriori model based on a posteriori information obtained from images and given by

$$\boldsymbol{r}=\boldsymbol{M}\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{n}}=\boldsymbol{d}\hat{\boldsymbol{\alpha}}_i+\boldsymbol{U}\hat{\boldsymbol{\gamma}}+\hat{\boldsymbol{n}} \tag{5}$$

where $\hat{\boldsymbol{\alpha}},\hat{a}_i$ and $\hat{\boldsymbol{\gamma}}$ are estimates of $\boldsymbol{\alpha},\alpha_i$ and $\boldsymbol{\gamma}$,respectively, based on the observed pixel $\boldsymbol{r}$ itself. Let $\boldsymbol{P}_{\boldsymbol{M}}=\boldsymbol{M}(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T$ be the signature space orthogonal projector, $<\boldsymbol{d}>$ and $<\boldsymbol{U}>$ be two non-orthogonal subspaces.

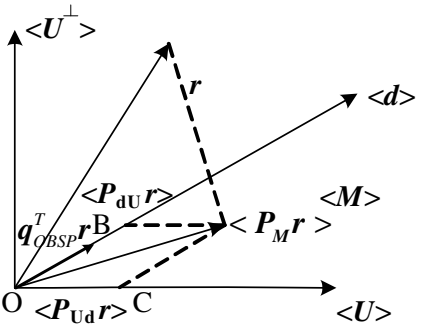


Fig. 1. Oblique Subspace Projection (OBSP)

The OBSP designates $< \boldsymbol{d} >$ as its range space $< \boldsymbol{U} >$ and as its null space. In other words, the OBSP maps an observed pixel into the target signature space, while annihilating all undesired signatures by mapping into the null space. Figure 1 shows the sketch map of the OBSP approach, where $\boldsymbol{P}_{\boldsymbol{M}}$ is decomposed as a sum of two oblique subspace projections

$$\boldsymbol{P}_{\boldsymbol{M}} = \boldsymbol{P}_{\boldsymbol{dU}} + \boldsymbol{P}_{\boldsymbol{Ud}} \tag{6}$$

with

$$\boldsymbol{P}_{\boldsymbol{dU}} = \boldsymbol{d}(\boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \boldsymbol{d})^{-1} \boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \tag{7}$$

$$\boldsymbol{P}_{\boldsymbol{Ud}} = \boldsymbol{U}(\boldsymbol{U}^T \boldsymbol{P}_{\boldsymbol{d}}^{\perp} \boldsymbol{U})^{-1} \boldsymbol{U}^T \boldsymbol{P}_{\boldsymbol{d}}^{\perp} \tag{8}$$

where $\boldsymbol{P}_{\boldsymbol{U}}^{\perp} = \boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}^T \boldsymbol{U})^{-1} \boldsymbol{U}^T$ denotes the background rejection operator. Particularly, $\boldsymbol{P}_{\boldsymbol{dU}} \boldsymbol{d} = \boldsymbol{d}$, $\boldsymbol{P}_{\boldsymbol{dU}} \boldsymbol{U} = 0$.

The OBSP operator is given by

$$\boldsymbol{q}_{\boldsymbol{OBSP}}^T = \boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{dU}} = \boldsymbol{d}^T \boldsymbol{d}(\boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \boldsymbol{d})^{-1} \boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \tag{9}$$

Applying (9) to posteriori model (5) results in

$$\boldsymbol{q}_{\boldsymbol{OBSP}}^T \boldsymbol{r} = \boldsymbol{d}^T \boldsymbol{d} \hat{\alpha}_1 \tag{10}$$

where $\boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{dU}} \hat{n} = 0$.

So, from (9) and (10), a normalized OBSP detection operator can be defined by

$$\boldsymbol{D}_{\boldsymbol{OBSP}}(\boldsymbol{r}) = \frac{\boldsymbol{q}_{\boldsymbol{OBSP}}^T \boldsymbol{r}}{\boldsymbol{d}^T \boldsymbol{d}} = (\boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \boldsymbol{d})^{-1} \boldsymbol{d}^T \boldsymbol{P}_{\boldsymbol{U}}^{\perp} \boldsymbol{r} \tag{11}$$

According to the analysis in [7], SSP and OBSP are essentially equivalent, in the sense that they both generate identical results.

3.2 OBSP in Feature Space and Its Kernel Version

A new mixture model can be defined by assuming that the input data has been implicitly mapped by a nonlinear function ϕ into a high-dimensional feature space $\mathcal{F}$. The mixture model is linear and additive in $\mathcal{F}$ and given by

$$\phi(\boldsymbol{r}) = \boldsymbol{m}_\phi \boldsymbol{\alpha}_\phi + \boldsymbol{n}_\phi \tag{12}$$

where $\boldsymbol{m}_\phi$ is a matrix whose columns are the endmember spectra in the feature space; $\boldsymbol{\alpha}_\phi$ is a coefficient vector that accounts for th abundances of each endmember spectrum in the feature space; $\boldsymbol{n}_\phi$ is an additive zero-mean Gaussian noise with covariance matrix $\sigma_\phi{}^2 \mathrm{I}_\phi$ (σ_ϕ^2: noise variance) and I_ϕ is the identity matrix in the feature space. The model (4) can also be rewritten as

$$\phi(\boldsymbol{r}) = \phi(\boldsymbol{d})\boldsymbol{\alpha}_{i\phi} + \boldsymbol{U}_\phi \boldsymbol{\alpha}_{U\phi} + \boldsymbol{n}_\phi \tag{13}$$

where $\phi(\boldsymbol{d})$ represent the spectral signature of the desired target with the corresponding abundance $\boldsymbol{\alpha}_{i\phi}$ in the feature space and the columns of $\boldsymbol{U}_\phi$ represent the undesired background signatures in the feature space.

According to the normalized operator of OBSP and the derivation of LSMM in the feature space, it is easy to obtain the OBSP operator in the feature space and the result of normalized OBSP for the pixel vector $\boldsymbol{r}$.

$$\boldsymbol{q}_{OBSP\phi}^T = \phi(\boldsymbol{d})^T \phi(\boldsymbol{d})[\phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp \phi(\boldsymbol{d})]^{-1} \phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp \tag{14}$$

$$\boldsymbol{D}_{OBSP\phi}(\boldsymbol{r}) = (\phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp \phi(\boldsymbol{d}))^{-1} \phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp \phi(\boldsymbol{r}) \tag{15}$$

where $\boldsymbol{P}_{d\phi U\phi} = \phi(\boldsymbol{d})[\phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp \phi(\boldsymbol{d})]^{-1} \phi(\boldsymbol{d})^T \boldsymbol{P}_{U\phi}^\perp$, $\boldsymbol{P}_{U_\phi}^\perp = \boldsymbol{I}_\phi - \boldsymbol{U}_\phi \boldsymbol{U}_\phi^\#$, $\boldsymbol{I}_\phi$ is a unit matrix in the feature space.

Due to it is difficulty to confirm an actual nonlinear mapping ϕ, equations (14)-(15) is not able to directly implement. However, it can be implemented by employing kernel functions using an efficient kernel trick without any knowledge of the actual nonlinear mapping ϕ. In the following, we show how to kernelize the OBSP in the feature space.

Suppose the rank of matrix $\boldsymbol{U}_\phi^T \boldsymbol{U}_\phi$ is N_b, $\lambda_i (i = 1, 2, \cdots, N_b)$ are the nonzero eigenvalues of matrix $\boldsymbol{U}_\phi^T \boldsymbol{U}_\phi$, $\boldsymbol{\Sigma}_\phi = diag(\lambda_1, \lambda_2, \cdots, \lambda_{N_b})$, and $\boldsymbol{A}_\phi$ is the eigenvector matrix of $\boldsymbol{U}_\phi^T \boldsymbol{U}_\phi$ corresponding to $\boldsymbol{\Sigma}_\phi$, then using the singular value decomposition theorem and its deduction [13], we get

$$\boldsymbol{U}_{\phi}^{T}\boldsymbol{U}_{\phi}=\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}\boldsymbol{A}_{\phi}^{T} \tag{16}$$

Therefore,

$$\boldsymbol{P}_{U_{\phi}}^{\perp}=\boldsymbol{I}_{\phi}-\boldsymbol{U}_{\phi}\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}^{-1}\boldsymbol{A}_{\phi}^{T}\boldsymbol{U}_{\phi}^{T} \tag{17}$$

Instituting equations (17) into equation (15) results in

$$\begin{aligned}\boldsymbol{D}_{OBSP\phi}(\boldsymbol{r})&=[\phi(\boldsymbol{d})^{T}\boldsymbol{P}_{U\phi}^{\perp}\phi(\boldsymbol{d})]^{-1}\phi(\boldsymbol{d})^{T}\boldsymbol{P}_{U\phi}^{\perp}\phi(\boldsymbol{r})\\&=\frac{\phi(\boldsymbol{d})^{T}\phi(\boldsymbol{r})-\phi(\boldsymbol{d})^{T}\boldsymbol{U}_{\phi}\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}^{-1}\boldsymbol{A}_{\phi}^{T}\boldsymbol{U}_{\phi}^{T}\phi(\boldsymbol{r})}{\phi(\boldsymbol{d})^{T}\phi(\boldsymbol{d})-\phi(\boldsymbol{d})^{T}\boldsymbol{U}_{\phi}\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}^{-1}\boldsymbol{A}_{\phi}^{T}\boldsymbol{U}_{\phi}^{T}\phi(\boldsymbol{d})}\end{aligned} \tag{18}$$

each element of vectors $\phi(\boldsymbol{d})^{T}\boldsymbol{U}_{\phi}$, $\boldsymbol{U}_{\phi}^{T}\phi(\boldsymbol{d})$ is also a kernel function, denote these vectors as $\boldsymbol{K}_{\boldsymbol{dU}}$ and $\boldsymbol{K}_{\boldsymbol{Ud}}$, $k_{\boldsymbol{dr}}=\phi(\boldsymbol{d})^{T}\phi(\boldsymbol{r})$ and $k_{\boldsymbol{dd}}=\phi(\boldsymbol{d})^{T}\phi(\boldsymbol{d})$ are kernel functions too. Now, equation (18) can be simply rewritten as

$$\boldsymbol{D}_{KOBSP}(\boldsymbol{r})=\frac{k(\boldsymbol{d},\boldsymbol{r})-\boldsymbol{K}(\boldsymbol{d},\boldsymbol{U})\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}^{-1}\boldsymbol{A}_{\phi}^{T}\boldsymbol{K}(\boldsymbol{U},\boldsymbol{r})}{k(\boldsymbol{d},\boldsymbol{d})-\boldsymbol{K}(\boldsymbol{d},\boldsymbol{U})\boldsymbol{A}_{\phi}\boldsymbol{\Sigma}_{\phi}^{-1}\boldsymbol{A}_{\phi}^{T}\boldsymbol{K}(\boldsymbol{U},\boldsymbol{d})} \tag{19}$$

Using equation (19), D_{KBOSP} can be calculated by kernel trick without any knowledge of the actual nonlinear mapping ϕ.

4 Experiments

4.1 Experiment Using Synthetic Hyperspectral Data

Firstly, we select three spectral signatures from the USGS digital spectral library [14], which are Carnallite, Ammonioalunite, and Biotite. The synthetic nonlinear mixture data was created using the Hapke Model described by Bo wu[15]. The datasets has 50 mixture pixels, and the angle of incidence is 30° and SNR is 30:1. The abundance fractions of the 50 synthetic mixture pixels are shown in Table 1.

Table 1. Assigning relative abundances of endmembers for datasets

	1-19	**20**	**21**	**…**	**29**	**30-50**
Carnallite	0.5	0.45	0.455	…	0.495	0.5
Ammonioaluni-te	0.5	0.45	0.455	…	0.495	0.5
Biotite	0	0.1	0.09	…	0.01	0

To evaluate the performance of the three algorithms, correlation coefficients, spectral angle distance (SAD) [16] are used to measure the similarity between the true endmember signature and its estimate, correlation coefficients(Cor) and root mean square error (RMSE) [17] are used to measure the similarity between the true abundances and the estimated abundances.

By several experiments, the kernel parameters of KOBSP was chosen as $a = 0.1$, b=0. 01, $c = 0.8$ for the datasets. Table 2 gives the three performance comparison between the abundances of target generated by OBSP and KOBSP based on Cor, AAD, and RMSE for the datasets. The detection result curve is shown in Figure 2.

From Figure 2, it is shown that the result obtained by KOBSP is very similarly with real abundance, and it also shows that the KOBSP obviously restrains the noise. The performance comparison in table 2 demonstrate that all the three performances of KOBSP are better than that of OBSP.

Table 2. The performance comparison between OBSP and KOBSP on the accuracy of abundance

	Cor	AAD	RMSE
OBSP	0.991711	0.12921	0.0495267
KOBSP	0.996843	0.0914265	0.0344829

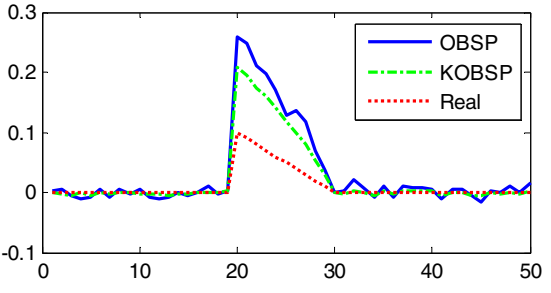


Fig. 2. The comparison of performance curves of OBSP, KOBSP and real abundance

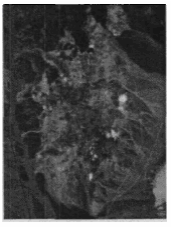
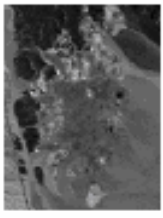
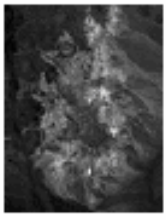


Fig. 3. Band 30 of the real hyperspectral scene

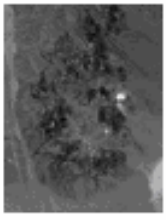


(a) pyrope



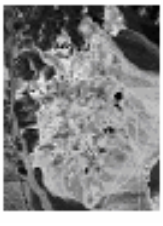
(b) alunite

(c) dersert varnish



(d) muscovite

Fig. 4. The detect results of OBSP for the four objects



(a) pyrope

(b) alunite

(c) dersert varnish

(d) muscovite

Fig. 5. The detect results of KOBSP for the four objects

4.2 Experiment for Real Image

In this section, we apply the two algorithms to real hyperspectral data captured by the AVIRIS over Cuprite, Nevada, shown in Figure 3, which has been widely used for remote sensing experiments. Our study is based on a subimage (250×190 pixels). To

improve the unmixing performance, we have removed the low SNR bands as well as the water-vapor absorption ban-ds (including bands 1-2, 104-113, 148-167, and 221-224) from the original 224 band data cube. Therefore, a total of 188 bands are used in the experiment[18].

In this experiment, we choose the RBF kernel as the kernel function of KOBSP. Its parameter $2\sigma^2$ is proved by experiments to be between 1.5 and 3.5. Figure 4 is the detect results of OBSP for the object pyrope, alunite, dersert varnish and muscovite respectively. Figure 5 is the results of KOBSP for the four objects with $2\sigma^2 = 2.52$. In Figure 4 and Figure 5, pure black denotes that the abundance of a certain material in this pixel is 0%, whereas pure white denotes 100%. As shown in these figures, each object is detected, but the results of OBSP are more blurry than that of KOBSP.

5 Conclusion

In this paper, we extended OBSP to its corresponding kernel version called KOBSP by using the idea of kernel-based learning theory. Experimental results based on simulated hyperspectral data and real hyperspectral imagery have indicated that KOBSP outperforms the conventional OBSP. Both OBSP and KOBSP have typically used the spectral information only, further work is to incorporate contextual information to KOBSP with various effective methods so as to obtain more perfect results.

Acknowledgments. This work is supported by Natural Science Basic Research Plan in Zhejiang Province of China Grant Y1100196 to X.R.Li.

References

1. Reed, I.S., Yu, X.: Addaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. IEEE Trans. Acountics. Speech and Signal Process 38(10), 1760–1770 (1990)
2. Harsanyi, J.C., Chang, C.-I.: Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Trans. on Geoscience and Remote Sensing 32(4), 779–785 (1994)
3. Scharf, L.L., Friedlander, B.: Matched subspace detectors. IEEE Trans. Signal Process. 42(8), 2146–2157 (1994)
4. Manolakis, D., Shaw, G., Keshava, N.: Comparative analysis of hyperspectral adaptive matched filter detectors. In: Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI, Orlando,Fla, USA. Proceedings of SPIE, vol. 4049, pp. 2–17 (April 2000)
5. Kraut, S., Scharf, L.L., McWhorter, T.: Adaptive subspace detectors. IEEE Trans. Signal Process. 49(1), 1–16 (2001)
6. Kraut, S., Scharf, L.L.: The CFAR adaptive subspace detectoris a scale-invariant GLRT. IEEE Transactions on SignalProcessing 47(9), 2538–2541 (1999)
7. Cheng-I, C., Xiaoli, Z., et al.: Least squares subspace projection approach to mixed pixel classification for hyperspectral images. IEEE Transactions on Geoscience and Remote Sensing 36(3), 898–912 (1998)

8. Tong, Q.X., Zhang, B., Zheng, L.F.: Hyperspectral remote sensing. Higher Education Press, Beijing (2006)
9. Kwon, H., Nasrabadi, N.M.: Kernel Orthogonal subspace projection for hyperspectral signal classifycation. IEEE Transactions on Geoscience and Remote Sensing 43(12), 2952–2962 (2005)
10. Capobianco, L., Camps-Valls, G., Garzelli, A.: Target detection with a contextual kernel orthogo nal subspace projection. In: SPIE International Symposium Remote Sensing XIV, Cardiff, UK (September 2008)
11. Zhao, L.X., Zhang, K., Li, X.R.: Kernel Signature Space Orthogonal Projection for Target Detection in Hyperspectral Imagery. Journal of Remote Sensing (in press)
12. John, S.T., Nello, C.: Kernel Methods for Pattern Analysis. China machine press (2005)
13. Bian, Z.Q., Zhang, X.G.: Pattern Recognition. Tsinghua University Press, Beijing (2000)
14. http://speclab.cr.usgs.gov
15. Wu, B., Zhang, L.P., Li, P.X.: Unmixing hyperspectral imagery based on support vector nonlinear approximating regression. Journal of Remote Sensing 10(3), 312–318 (2006)
16. Miao, L., Qi, H.: Endmember Extraction From Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization. IEEE Transactions on Geoscience and Remote Sensing 45(3), 765–777 (2007)
17. Sen, J.: Unsupervised hyperspectral unming theory and techniques. Ph.D Dissertation, Zhejiang University (2007)
18. Swayze, G.: The hydrothermal and structural history of the Cuprite Mining District, Southwestern Nevada: An integra-ted geological and geophysical approach, Ph.D. dissertation, University of Colorado, Boulder (1997), `http://speclab.cr.usgs.gov` and `ftp://ftpext.cr.usgs.gov/pub/cr/co/denver/speclab/pub/cuprit/gregg.thesis.images/mineralzone.thesis.gif`

Text-Independent Voice Conversion Based on Kernel Eigenvoice

Yanping Li[1], Linghua Zhang[1], and Hui Ding[2]

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
[2] College of Mathematics and Information Engineering, Jiaxing University, Jiaxing, China
{liyp,zhanglh}@njupt.edu.cn, dh_jxxy@yahoo.com.cn

Abstract. Almost of the current spectral conversion methods required parallel corpus containing the same utterances from source and target speakers, which was often inconvenient and sometimes hard to fulfill. This paper proposed a novel algorithm for text-independent voice conversion, which can relax the parallel constraint. The proposed algorithm was based on speaker adaptation technique of kernel eigenvoice, adapting the conversion parameters derived for the pre-stored pairs of speakers to a desired pair, for which only a nonparallel corpus was available. Objective evaluation results demonstrated that the proposed kernel eigenvoice algorithm can effectively improve converted spectral similarity in a text-independent manner.

Keywords: voice conversion, Gaussian mixture model, text independent, kernel eigenvoice, adaptation.

1 Introduction

The aim of voice conversion (VC) is to change the speaker individuality [1] [2]. Many approaches to VC have been studied, which include vector quantization(VQ)[3], codebook mapping, sentence hidden Markov Model (HMM) and Gaussian mixture model (GMM), radial basis function networks (ANN)[1,2,3].

However, these traditional voice conversion techniques are text-dependent, i.e., they require parallel training data which must be further aligned using e.g. dynamic time warping. From the viewpoint of practical applications, the requirement of having parallel speech databases is rather inconvenient and sometimes hard to fulfill. Moreover, in some cases it may be even impossible to obtain parallel speech corpora. This is the case e.g. in cross-lingual voice conversion where the source and the target speakers speak in different languages [4]. Due to these reasons, it is important to develop text-independent voice conversion techniques that can be trained using non-parallel databases.

In this paper, we would like to apply kernel methods to improve the performance of eigenvoice-based methods for text-independent voice conversion. The proposed algorithm will exploit possible nonlinearity in the speaker supervector space, and investigate the use of nonlinear PCA to find better eigenvoices.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 432–439, 2010.

2 Conventional GMM-Based Mapping

2.1 Parallel Training

Let x_t and y_t be D-dimensional source and target feature vectors at frame t, respectively. The joint probability density of the source and target feature vectors is modeled by a GMM as follows [5]:

$$P(z_t \left| \lambda^{(z)} \right.) = \sum_{m=1}^{M} w_m N(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \tag{1}$$

Where z_t is a joint vector $\left[x_t^T, y_t^T \right]^T$. The mixture component index is m. The weight of the m^{th} mixture component is w_m. The mean vector $\mu_m^{(z)}$ and the covariance matrix $\Sigma_m^{(z)}$ of the m^{th} mixture component are written as:

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{2}$$

Where $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vector of the m^{th} mixture component for the source and that for the target. The matrices $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the covariance matrix of the m^{th} mixture component. The matrices $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the cross covariance matrices.

2.2 Mapping Function

The conditional probability density of y_t, given x_t, is also represented as a GMM as follows:

$$P(y_t \mid x_t, \lambda^{(z)}) = \sum_{m=1}^{M} P(m \mid x_t, \lambda^{(z)}) P(y_t \mid x_t, m, \lambda^{(z)}) \tag{3}$$

$$P(m \mid x_t, \lambda^{(z)}) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})} \tag{4}$$

$$P(y_t \mid x_t, m, \lambda^{(z)}) = N(y_t; E_{m,t}^{(y)}, D_m^{(y)}) \tag{5}$$

The mean vector $E_{m,t}^{(y)}$ and the covariance matrix $D_m^{(y)}$ of the m^{th} conditional probability distribution are written as:

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (x_t - \mu_m^{(x)}) \tag{6}$$

$$D_m^{(y)} = \Sigma_m^{(yy)} - \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} \Sigma_m^{(xy)} \tag{7}$$

The conversion is performed on the basis of the minimum mean-square error (MMSE) as follows:

$$\begin{aligned}\hat{y}_t &= E[y_t \mid x_t] = \int P(y_t \mid x_t, \lambda^{(z)}) y_t dy_t \\ &= \int \sum_{m=1}^{M} P(m \mid x_t, \lambda^{(z)}) P(y_t \mid x_t, m, \lambda^{(z)}) y_t dy_t = \sum_{m=1}^{M} P(m \mid x_t, \lambda^{(z)}) E_{m,t}^{(y)}\end{aligned} \tag{8}$$

Where $E[\cdot]$ means the expectation and $\hat{y}_t$ is the converted target feature vector. The converted feature vector is defined as the weighted sum of the conditional mean vectors, where the posterior probabilities in (4).

3 Proposed Text-Independent Kernel Eigenvoice Conversion

EVC has three main processes: training, adaptation and conversion. In the training process, an eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets consisting of the specified source speaker and many pre-stored target speakers. In the adaptation process, the trained EV-GMM is capable of being flexibly adapted to a new target speaker using only a few arbitrary utterances of the target speaker. A few adaptive parameters are estimated in a completely text-independent manner. In the conversion process, arbitrary utterances of the source speaker's voice are converted into those of the new target speaker's voice using adapted EV-GMM. The whole process of EV-GMM suggests that the acoustic feature correlation between pre-stored speaker pair is effectively used as a prior knowledge for building the correlation of new speaker pair. In the following subsections, the EV-GMM's structure and the proposed kernel eigenvoice conversion (KEVC) processes are described in detail.

3.1 Eigenvoice Gaussian Mixture Model

EV-GMM represents the joint probability density in the same manner as the conventional GMM shown in (1), except for a definition of the target mean vector [5] written as:

$$\mu_m^{(y)} = B_m \mathbf{w} + b_m^{(0)} \tag{9}$$

Where $b_m^{(0)}$ is a bias vector for the m^{th} mixture. The matrix $B_m = [b_m(1), \cdots b_m(J)]$ consists of basis vectors $b_m(j)$ for the m^{th} mixture. The number of basis vectors is J. The target speaker individuality is controlled with the J-dimensional weight vector $\mathbf{w} = [w(1), \cdots w(J)]^T$. Consequently, the EV-GMM has a parameter set $\lambda^{(\mathrm{EV})}$ consisting of the single weight vector and parameters for individual mixtures such as the mixture weights, the source mean vectors, the bias and basis vectors, and the covariance matrices.

3.2 Training of EV-GMM Based on Principal Component Analysis

First, a target-speaker-independent GMM (TI-GMM) $\lambda^{(0)}$ is trained with multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers as follows:

$$\hat{\lambda}^{(0)} = \arg\max_{\lambda^{(0)}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(x_t, y_t^{(s)} \mid \lambda^{(0)}) \tag{10}$$

Then, using only the parallel data set for the s^{th} pre-stored target speaker, the s^{th} target-speaker-dependent GMM (TD-GMM) $\lambda^{(s)}$ is trained by only updating target mean vectors $\mu_m^{(y)}$ of $\lambda^{(0)}$ based on maximum likelihood (ML) estimate as follows:

$$\hat{\lambda}^{(s)} = \arg\max_{\mu_m^{(y)}} \prod_{t=1}^{T_s} P(x_t, y_t^{(s)} \mid \lambda^{(0)}) \tag{11}$$

After training the TD-GMMs for all pre-stored target speakers, a $2 \times D \times M$ dimensional supervector $SV^{(s)} = [\mu_{1s}^{(y)^T}, \mu_{2s}^{(y)^T}, \cdots \mu_{Ms}^{(y)^T}]^T$ is constructed for each pre-stored target speaker by concatenating the updated target mean vectors $\{\mu_{1s}^{(y)}, \mu_{2s}^{(y)}, \cdots \mu_{Ms}^{(y)}\}$ of $\lambda^{(s)}$. Finally, bias vector $b_m^{(0)}$ and basis vectors B_m are extracted by performing principal component analysis (PCA) for the supervectors $\{SV^{(1)}, SV^{(2)}, \cdots SV^{(S)}\}$ for all pre-stored speakers, where S denotes the number of pre-stored target speakers. Each supervector is approximated as follows:

$$SV^{(s)} \approx [B_1^T, B_2^T, \cdots B_M^T]^T \mathbf{w}^{(s)} + [b_1^{(0)^T}, b_2^{(0)^T}, \cdots b_M^{(0)^T}]^T \tag{12}$$

$$b_m^{(0)} = \frac{1}{S} \sum_{s=1}^{S} \mu_{ms}^{(Y)} \tag{13}$$

Where $\mathbf{w^{(s)}}$ consists of $J (< S << 2DM)$ principle components for the s^{th} pre-stored target speaker.

3.3 Training of the Proposed KEV-GMM Based on Kernel Principal Component Analysis

As mentioned above, in traditional EV-GMM algorithm, the eigenvoice is found by linear PCA. Then a new speaker is represented as a linear combination of a few most important eigenvoices, and the eigenvoice weights are usually estimated by maximizing the likelihood of the adaptation data. In this paper, we will exploit possible nonlinearity in the speaker supervector space, and investigate the use of nonlinear PCA to find the eigenvoices by kernel methods. We will call the proposed method kernel eigenvoice conversion (KEVC). In principle, since the KEV conversion method is a nonlinear generalization of the EV conversion method, the former should be more powerful than the latter, and KEV conversion is expected to give better performance.

Let $k(\cdot,\cdot)$ be the kernel with an associated mapping φ which maps a pattern $\mathbf{x} \in \mathbf{R}^{d_2}$ (a speaker supervector in the eigenvoice approach) in the input space to $\varphi(\mathbf{x}) \in \mathbf{R}^{d_3}$ in the high-dimensional kernel- induced feature space. Given a set of S patterns $\{\mathbf{x_1}, \cdots \mathbf{x}_S\}$, their φ-mapped feature vectors are $\{\varphi(\mathbf{x_1}), \cdots \varphi(\mathbf{x}_S)\}$. The S mapped patterns are first centered in the feature space by finding the mean of the

feature vectors $\overline{\varphi} = (1/S)\sum_{i=1}^{S}\varphi(\mathbf{x_i})$. Let the "centered" map be $\tilde{\varphi}$ so that $\tilde{\varphi}(\mathbf{x}) = \varphi(\mathbf{x}) - \overline{\varphi}$. In addition, let $\mathbf{K} = [k_{ij}]$ be the kernel matrix with:

$$k_{ij} = k(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i})^T \varphi(\mathbf{x_j}) \tag{14}$$

And $\tilde{\mathbf{K}}$ be the centered version of $\mathbf{K}$ with $\tilde{k}_{ij} = \tilde{\varphi}(\mathbf{x_i})^T \tilde{\varphi}(\mathbf{x_j})$. Notice that $\tilde{\mathbf{K}}$ is related to $\mathbf{K}$ by $\tilde{\mathbf{K}} = \mathbf{HKH}$, where $\mathbf{H} = \mathbf{I} - (\mathbf{1}/\mathbf{N})\mathbf{1}\mathbf{1}^{\mathbf{T}}$ is the centering matrix, $\mathbf{I}$ is the $S \times S$ identity matrix, and $\mathbf{1} = [1,1,\cdots 1]^T$ is an S -dimensional vector. To perform kernel PCA, instead of directly working on the covariance matrix in the feature space, one may carry out eigen-decomposition on the centered kernel matrix $\tilde{\mathbf{K}}$ as $\tilde{\mathbf{K}} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U} = [\alpha_1, \cdots \alpha_S]$ with $\alpha_i = [\alpha_{i1}, \cdots \alpha_{iS}]^T$, and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_S)$. The j^{th} orthonormal eigenvector of the covariance matrix [13] in the feature space is then given as:

$$\mathbf{V}_j = \sum_{i=1}^{S} \frac{\alpha_{ji}}{\sqrt{\lambda_j}} \tilde{\varphi}(\mathbf{x_i}) \tag{15}$$

For the s^{th} supervector, $SV^{(s)} = [\mu_{1s}^{(y)^T}, \mu_{2s}^{(y)^T}, \cdots \mu_{Ms}^{(y)^T}]^T$, we map each constituent $\mu_{ms}^{(y)}$ via a separate kernel $k_m(\cdot,\cdot)$ to $\varphi_m(\mu_{ms}^{(y)})$ and construct $\varphi(SV^{(s)}) = [\varphi_1(\mu_{1s}^{(y)^T}), \cdots \varphi_M(\mu_{Ms}^{(y)^T})]^T$. Let the centered supervector of a new target speaker found by kernel eigenvoice method in the feature space be $\tilde{\varphi}^{(\mathrm{kev})}(s)$, which is assumed to be a linear combination of the first J eigenvectors with the largest eigenvalues found by kernel PCA . That is

$$\tilde{\varphi}^{(\mathrm{kev})}(s) = \sum_{j=1}^{J} w_j \mathbf{V}_j = \sum_{j=1}^{J}\sum_{i=1}^{S} \frac{w_j \alpha_{ji}}{\sqrt{\lambda_j}} \tilde{\varphi}(\mathbf{x_i}) \tag{16}$$

Its m^{th} constituent is then given by:

$$\tilde{\varphi}_m^{(\mathrm{kev})}(s_m) = \sum_{j=1}^{J}\sum_{i=1}^{S} \frac{w_j \alpha_{ji}}{\sqrt{\lambda_j}} \tilde{\varphi}_m(\mathbf{x_{im}}) \tag{17}$$

3.4 Unsupervised Adaptation of Trained KEV-GMM and Conversion

In adaptation stage, given the new desired target speaker non-parallel data $y = \{y_1, \cdots y_T\}$, the eigenvoice weights are usually estimated by maximizing the likelihood of Y . Notice that only the desired target speaker data is used for adaptation, then non-parallel data is qualify, furthermore, different language data is expected to convert. Mathematically, we can find $\mathbf{w}$ by maximizing the following auxiliary function [6]:

$$Q(\mathbf{w}) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_t(m)\left\|y_t - s_m(\mathbf{w})\right\|_{C_r^{-1}}^2 \tag{18}$$

In this paper, we apply the generalized EM algorithm [7] to find the optimal weight $\hat{\mathbf{w}}$ for the desired target speaker by solving the nonlinear equations.

3.5 Conversion with Adapted KEV-GMM

In the conversion stage, given the source speaker feature vectors $X = \{x_t, t = 1, \cdots, T\}$, the converted target feature vectors $\hat{y}$ are obtained in the same manner as traditional GMM, but the mean vector $E_{m,t}^{(y)}$ in (6) is modified as:

$$E_{m,t}^{(y)} = B_m\hat{\mathbf{w}} + b_m^{(0)} + \Sigma_m^{(yx)}\Sigma_m^{(xx)^{-1}}(x_t - \mu_m^{(x)}) \tag{19}$$

4 Experimental Results and Discussion

The speech data used in voice conversion is collected with 22.05 kHz sampling rate and 16-bit coding quantization. It consists of 60 speakers (28 females and 32 males). The whole database is recorded in Chinese Mandarin, including three parts. In digit session, each speaker pronounces ten sentences about 4 connected digits, the duration of each connected digits is about 6 seconds. In fixed session, each speaker reads 'the wind and the sun' in Chinese edition, which is about 1minute. In random session, every speaker pronounces ten sentences about routine matters, the duration of which is varying between 8~10 seconds. The speech data was weighted by Hamming window and divided into frames of 512 samples with a frame overlapping of 50% and pre-emphasized with coefficient 0.98. We use 24-dimensional mel-cepstrum as a spectral feature, which is extracted from smoothed spectrum analyzed by STRAIGHT. We evaluate mel-cepstral distortion between the target and converted features. Mel-cepstral distortion [5] is calculated as:

$$\text{Mel-CD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(m_d^t - m_d^c)} \tag{20}$$

Where m_d^t and m_d^c represent the d^{th} dimensional component of target mel-cepstrum and that of converted mel-cepstrum, respectively.

The purpose of this experiment is to evaluate the text-independent conversion performance using eigenvoice method. For constructing non-parallel conversion based on EV-GMM and KEV-GMM, we choose 50 speakers as pre-stored target speaker and 1 female speaker as source speaker (marked as F). The fixed content sessions of source speaker and pre-stored speakers are used for parallel training and the number of mixture components is $M = 256$. We choose 4 residual speakers (2 females and 2 males, which is marked as F1, F2 and M1, M2) as desired target speaker respectively. Digit session of desired target speaker is used as adaptation data, which is different from the source training data. In conversion stage, we choose random sentences as testing data. For constructing baseline system of traditional parallel GMM,

we use fixed session of source and desired target speaker as training data and random sentences as testing data. The number of mixture components for baseline system is $M = 128$. The contrast of performance for parallel and non-parallel conversion is showed in Fig.1.

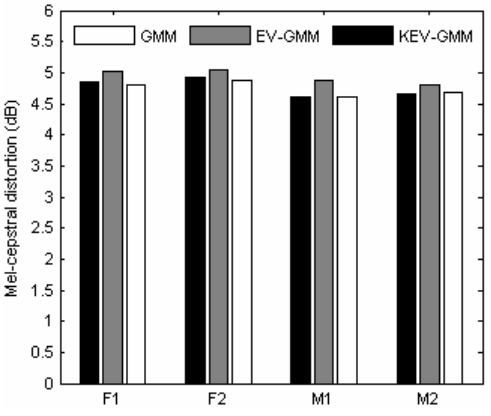


Fig. 1. The contrast of performance for parallel and non-parallel conversion methods

In Fig. 1, "F1" in the x-axis means the conversion from female source speaker F to target speaker female F1. From the analysis of four conversion results, we can draw two conclusions as follows:

The performance of baseline system based on GMM is superior to other two methods more or less, we think proximate cause of major difference is that the training and testing data is parallel based on GMM, while the training data for source speaker and desired speaker is non-parallel based on EV-GMM and KEV-GMM, there is an adaptation process and only the mean vector is adjusted. However, the two algorithms of EV-GMM and KEV-GMM have incomparable advantage of allowing nonparallel corpus.

The proposed KEV-GMM algorithm performs quite favorably, and the distortion is lower than EV-GMM, which demonstrated that kernel PCA can effectively analyze the nonlinear information from the supervectors and find better eigenvoices for adaptation. Further, the performance of KEV-GMM is somewhat inferior to baseline system of GMM, but the difference is not so large, in other words, the proposed algorithm can achieve comparable performance in nonparallel manner with baseline system using parallel data.

5 Conclusions and Future Works

A new approach to the concept of text-independent voice conversion has been developed in this study. The proposed algorithm considers the nonlinearity in the speaker supervector space, and investigates the use of nonlinear PCA to find the eigenvoices by kernel methods. Objective evaluate results indicated that the proposed

algorithm was a promising technique for text-independent voice conversion. The subjective listening test and cross-language conversion based on the proposed algorithm are the next work.

Acknowledgments. This work is supported by National Natural Science Foundation of China (60872105); Scientific Research Foundation for Young Teachers in Nanjing University of Posts and Telecommunications (NY209004); Zhejiang Natural Science Foundation of China (Y1090649); Foundation of Zhejiang Educational Committee (Y200805349).

References

1. Moulines, E., Sagisaka, Y.: Voice conversion: state of the art and perspectives. Speech Communication 16(2), 125–126 (1995)
2. Stylianou, Y.: Voice Transformation: a survey. In: IEEE International Conference, pp. 3585–3588. IEEE Press, Taibei (2009)
3. Kain, A.: High resolution voice conversion. Doctoral Thesis, Portland, Oregon: OGI School of Science and Engineering, Oregon Health and Science University (2001)
4. Sündermann, D., Bonafonte, A., Ney, H., et al.: A first step towards text-independent voice conversion. In: International Conference on Spoken Language Processing, Jeju Island, pp. 1173–1176 (2004)
5. Ohtani, Y.: Techniques for improving voice conversion based on eigenvoices. Doctoral Thesis, Nara Institute of Science and Technology (2010)
6. Brian, M., Kwok, T., Ho, S.: Kernel eigenvoice speaker adaptation. IEEE Transactions on Speech and Audio Processing 13(5), 984–992 (2005)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39(1), 1–38 (1977)

An Effective Method for SAR Automatic Target Recognition

Ying Li and Hongli Gong

School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi, 710129, China
lybyp@nwpu.edu.cn, gogoholy@163.com

Abstract. Since synthetic aperture radar (SAR) images are very sensitive to the pose variation of targets, SAR automatic target recognition (ATR) is a well-known very challenging problem. This paper introduces an effective method for SAR ATR by using a combination of kernel singular value decomposition (KSVD) and principal component analysis (PCA) for feature extraction and the nearest neighbor classifier (NNC) for classification. Experiments are carried out on the Moving and Stationary Target Acquisition and Recognition (MSTAR) public database to evaluate the performance of the proposed method in comparison with the traditional PCA, singular value decomposition (SVD), kernel PCA (KPCA) and KSVD. The results demonstrate that the proposed method performs much better than the other methods with a right recognition rate up to 95.75%.

Keywords: synthetic aperture radar (SAR) images; automatic target recognition (ATR); kernel singular value decomposition (KSVD); principal component analysis (PCA); feature extraction.

1 Introduction

Synthetic aperture radar (SAR) has been widely used in military and civil fields for its all weather and all time work capability. Over the last decades, researchers have drawn much attention on the automatic target recognition (ATR) based on SAR images.

Many template matching based methods which are popular nowadays have already been used in SAR ATR [1]-[6]. Since the recognition performance lies largely on the effectiveness of feature extraction, it is important to find an effective feature extract method. As a classical method for feature extraction, principal component analysis (PCA) and its variations have been used in SAR ATR for the purpose of data compression and redundancy reduction [2][3][6]. Although good recognition results are reported in the public papers, the PCA features are very sensitive to the target's azimuth angle. Therefore, the kernel based methods are used to extract the nonlinear features of images. Han et al. [1] proposed a SAR ATR algorithm using kernel PCA (KPCA) as feature extractor and support vector machine (SVM) as classifier, which could reach a right recognition rate of 91.5% on SAR images with no target azimuth information.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 440–447, 2010.

In this paper, an effective method for SAR ATR is proposed using a combination of kernel singular value decomposition (KSVD) and PCA for feature extraction and the concise and quick nearest neighbor classifier (NNC) for classification. The performance of the proposed method is evaluated on the Moving and Stationary Target Acquisition and Recognition (MSTAR) public database compared to the PCA, singular value decomposition (SVD), KPCA and KSVD methods. The experimental results show the effectiveness of the proposed method with a right recognition rate of 95.75% even if no target azimuth information is available.

2 Kernel SVD (KSVD)

2.1 SVD

Due to the advantages of stability, position invariance, rotation invariance, transposition invariance, mirror transformation invariance and so on, SVD can be regarded as an effective algebraic characteristic of images [7][8].

Firstly, create a data matrix $a=[x_1...x_n]\in R^{m*n}$ from image set, where $x_i \in R^m$ denotes the i-th input image. Without loss of generality, we assume that $m>n$. Given the SVD of a: $a=U\sum V^T$, we can compute the rank-r singular value factorization of a by

$$a^r = U^r \sum^r \left(V^r\right)^T . \tag{1}$$

where $r<n$, U^r, V^r are the first r columns of U, V, respectively, and $\sum^r = diag(\sigma_1,...,\sigma_r)$ with $\sigma_1 \geq ... \geq \sigma_r > 0$.

2.2 KSVD

Since SVD is a linear algorithm, nonlinear features can't be extracted effectively, but they exist widely in practical problems. Thus, nonlinear feature spaces may be more suitable to describe SAR images which are known to lying on highly complex manifolds. For this reason, similar to SVM, KSVD [9][10] takes advantage of the kernel trick to nonlinearly map the input data to a higher dimensional space using the function $\phi: R^m \to \mathrm{F}$, and perform the SVD in F. The map ϕ is induced by a kernel function $k(.,.)$ that evaluates the inner product between mapped input data in feature space.

We transform a into $A=[\phi(x_1)...\phi(x_n)]$ using ϕ assuming $\sum_{i=1}^{n}\phi(x_i)=0$. Consider $M=A^TA$ and its eigenvalue decomposition $M=Q\Delta Q^T$. Matrix M contains inner products between the columns of A and can be computed via the kernel function without having to evaluate $\phi(x_i)$. Then the rank-r singular value factorization of A is

$$A^r = \left[AQ^r\left(\Delta^r\right)^{-\frac{1}{2}}\right]\left[\left(\Delta^r\right)^{\frac{1}{2}}\right]\left[\left(Q^r\right)^T\right] \equiv U^r\sum^r\left(V^r\right)^T . \tag{2}$$

with Q^r being the first r columns of Q and Δ^r being the top left $r \times r$ square matrix of Δ. M is positive semi-definite if it is constructed using a Mercer kernel.

From formula (2) we can see that U^r is expanded by A in F

$$U^r = AQ^r \left(\Delta^r\right)^{-\frac{1}{2}} = A\alpha \cdot \tag{3}$$

Thus, the mapping of B in F to U^r is

$$\left(U^r\right)^T \cdot B = \alpha^T A^T B \ . \tag{4}$$

According to the Mercer theorem, we replace the inner products in F by the kernel function

$$K\left(A,B\right) = A^T B \ . \tag{5}$$

Therefore, formula (4) can be changed into

$$\left(U^r\right)^T \cdot B = \alpha^T K\left(A,B\right) \ . \tag{6}$$

It should be noted that we assume $\sum_{i=1}^{n} \phi\left(x_i\right) = 0$ at the start of this part. If the samples in F do not meet this condition, they should be replaced by

$$\tilde{\phi}\left(x_i\right) = \phi\left(x_i\right) - \frac{1}{n}\sum_{i=1}^{n} \phi\left(x_i\right) \ . \tag{7}$$

while other processes are the same as above. At the same time, the K in formula (5) changes into

$$\tilde{K} = K - l_n K - K l_n + l_n K l_n \ . \tag{8}$$

with l_n being the $n \times n$ matrix with all the elements being $1/n$.

Commonly used kernel functions are as follows.

- Polynomial kernel

$$K\left(x,x'\right) = \left\langle x,x' \right\rangle^d \ . \tag{9}$$

  with $x, x' \in R^m$ and d being integer, to avoid appearing zeros in kernel function we use $K\left(x,x'\right) = \left\langle x,x'+1 \right\rangle^d$ to replace (9).

- Gaussian radial basis kernel

$$K\left(x,x'\right) = exp\left(-\frac{\left\|x-x'\right\|^2}{2\delta^2}\right) \cdot \tag{10}$$

  with δ being the width of kernel.

- Index of radial basis kernel

$$K(x,x') = exp\left(-\frac{\|x-x'\|}{2\delta^2}\right) . \tag{11}$$

- Multi-layer perception kernel (sigmoid kernel)

$$K(x,x') = \tanh\left(p\langle x,x'\rangle + c\right) . \tag{12}$$

with p being the gain coefficient and c the threshold coefficient.

3 PCA

The central idea of PCA [1] is to reduce the dimensionality of a data set in which there are a large number of interrelated variables. Assume $X = \{x_1, x_2, ..., x_M\}$, where $x_j \in R^N$ ($j = 1, 2, ..., M$) is a column vector, M is the number of images in the training set. Then the maximum likelihood estimation of the covariance matrix among the M samples is

$$C_x = \frac{1}{M}\sum_{j=1}^{M} d_j d_j^T . \tag{13}$$

where $d_j = x_j - m$, $m = \frac{1}{M}\sum_{j=1}^{M} x_j$.

Apply eigenvalue decomposition on C_x

$$\lambda_i v_i = C_x v_i . \tag{14}$$

where λ_i is the eigenvalue of C_x with $i = 1, 2, ..., N$ and v_i is the corresponding eigenvector. The eigenvectors corresponding to n largest eigenvalues are used to form the mapping matrix. Thus, using the mapping matrix, we can transform both the training and testing data into an n-dimensional feature space. Usually, we choose n according to the contribution rate η as the following formula

$$\eta = \sum_{j=1}^{n} \lambda_j \Big/ \sum_{j=1}^{N} \lambda_j . \tag{15}$$

where η is a constant number between 0 and 1, usually set $\eta \geq 85\%$.

4 Feature Extraction Based on KSVD and PCA

To make the extracted features stable and remove the redundancy among the features, we combine KSVD and PCA to extract the features of SAR images. The concrete steps are as follows.

- Feature extraction on training set:

1) Transform each image in the training set into a column vector and standardize them.

2) Apply KSVD on the standardized matrix to obtain the first feature $KSVDFeature$ and mapping matrix $KSVDMap$.
3) Further carry out PCA on $KSVDFeature$ to get the last feature $TRAINFeature$ and mapping matrix $PCAMap$.

- Feature extraction on testing set:

1) Transform each image in the test set into a column vector and standardize them.
2) Use kernel function to map the standardized matrix to the high dimensional space and obtain the centralized kernel matrix $TESTMidKernel$ for testing set.
3) From the training process we can get the last mapping matrix $Map = KSVDMap * PCAMap$ and then the feature of testing set is $TESTFeature = Map^T * TESTMidKernel$.

5 Experimental Results and Analysis

MSTAR public dataset, including T72, BMP2 and BTR70, is used to compare the performance of the proposed method with traditional PCA, SVD, KPCA and KSVD methods. The data is actual measurement of ground stationary targets, and it is gathered by spotlight SAR with $0.3\times0.3m$ resolution. In the dataset, all images are 128 by 128 pixels, and the azimuth coverage for each type targets is 0^0 to 360^0 with 1^0 interval. The training and testing targets are collected at a depression angle of 17^0 and 15^0, respectively. The number of the three types of training targets is 232, 233 and 233, and the testing targets 196, 196 and 196, respectively. A part of the optical images of three types of tank are shown in Fig. 1, and SAR images in Fig. 2.

Fig. 1. A part of the optical images of three types of tank

We carry out the experiments using a combination of NNC and the different feature extractors, including the proposed KSVD and PCA, PCA, SVD, KPCA, KSVD, for recognition. For all these feature extraction methods, if there is a need, the feature dimension is set as 20, and the Gaussian radial basis kernel function with $\delta = 100$ is adopted. Shown in Fig. 3 is the flow chart for recognition of the proposed method, where solid lines show the training process, and dash lines show the testing process.

Table 1 shows the recognition results on SAR images. We can see that the performance of the proposed KSVD and PCA based method is better than that of the PCA, SVD, KPCA and KSVD based methods when no target azimuth information is known, which indicates that the proposed method can greatly reduce the sensitivity to

target azimuth for SAR images. This is mainly because we make a more efficient target feature extraction using a combination of KSVD and PCA. In addition, the NNC we used is concise and quick with lower computation times.

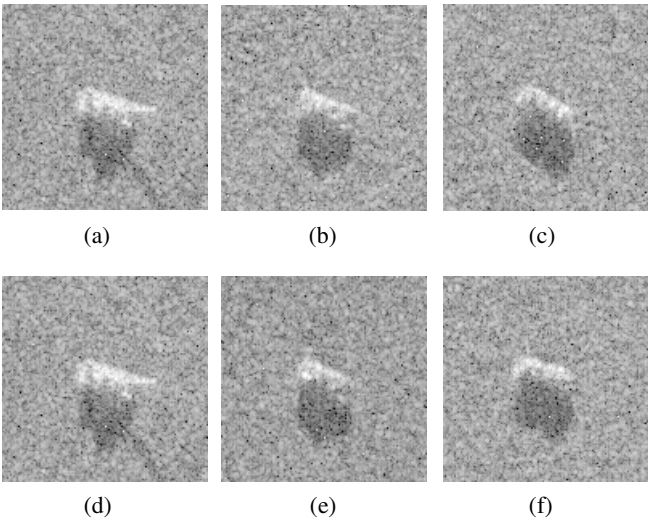


Fig. 2. A part of MSTAR dataset. (a)~(c) are the BMP2, BTR70, T72 targets in the training set respectively, (d)~(f) are the corresponding targets in the testing set.

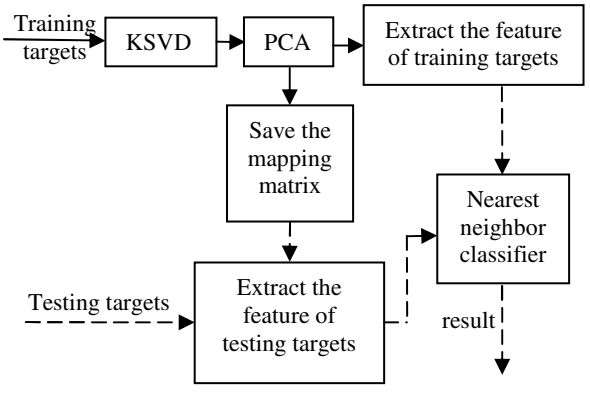


Fig. 3. The flow chart for recognition

Table 1. Comparison of different ATR methods without target azimuth information (%)

Different methods for feature extraction combined with NNC				
PCA	SVD	KPCA	KSVD	KSVD+PCA
64.8	88.27	87.42	90.8	95.75

6 Conclusion

In this paper, an effective method for SAR ATR based on the combination of KSVD and PCA is proposed. First, the original targets are nonlinearly mapped to a high dimensional subspace using the kernel function. Then we use linear SVD in the subspace to get the nonlinear structural features. We further apply PCA to get the final classification features. At last, NNC is used for classification. In this way, we make full use of the advantages of KSVD and PCA, which makes the design of classifier more concise and effective. The experimental results on MSTAR dataset indicate that the performance of the proposed method is superior to that of PCA, SVD, KPCA and KSVD, with a higher right recognition rate without target azimuth information. Moreover, the NNC we used is very efficient for its simplicity. Therefore, the method proposed in this paper is effective for SAR ATR.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 60873086), the Specialized Research Found for the Doctoral Program of Higher Education (No. 20070699013), and the Aeronautical Science Foundation of China (No. 20085153013).

References

1. Han, P., Wu, R.B., Wang, Z.H., Wang, Y.H.: SAR Automatic Target Recognition based on KPCA Criterion. Journal of electronics and information technology 25, 1297–1301 (2003)
2. Yang, Y.N., Qiu, Y.X., Lu, C.: Automatic Target Classification Experiments on the MSTAR SAR Images. In: 6th International Conference on SNPD/SAWN, pp. 2–7. IEEE Computer Society, Washington (2005)
3. Duan, R., Chen, L., Man, H.: Adaptive SAR ATR in likelihood space. In: Algorithms for Synthetic Aperture Radar Imagery XII, Orlando, FL, USA. Proceedings of SPIE, vol. 5808, pp. 294–305 (2005)
4. Mishra, A.K., Mulgrew, B.: Bistatic SAR ATR using PCA-based features. In: Automatic target recognition XVI, Kissimmee, Florida, USA. Proceedings of SPIE, vol. 6234, pp. 62340Ul–62340U9 (2006)
5. Chen, Y., Blasch, E., Chen, H.M., Qian, T., Chen, G.S.: Experimental Feature-Based SAR ATR Performance Evaluation under Different Operational Conditions. In: Signal processing, sensor fusion, and target recognition XVII, Orlando, Florida, USA. Proceedings of SPIE, vol. 6968, pp. 69680F.1–69680F.12 (2008)
6. Wang, S.X., He, Z.G.: The Fast Target Recognition Approach Based on PCA Features for SAR Images. Journal of National University of Defense Technology 30, 137–141 (2008)
7. Zhou, D.L., Gao, W., Zhao, D.B.: Face Recognition Based on Singular Value Decomposition and Discriminant KL Projection. Journal of Software 14, 783–789 (2003)

8. Liu, Z.H., Jin, Z., Lai, Z.H., Huang, C.B., Wan, M.H.: Face Recognition Based on Wavelet Transform, Singular Value Decomposition and Kernel Principal Component Analysis. In: Chinese Conference on Pattern Recognition 2008 (CCPR 2008), pp. 1–5. IEEE Press, Beijing (2008)
9. Chin, T., Schindler, K., Suter, D.: Incremental Kernel SVD for Face Recognition with Image Sets. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR), pp. 461–466. IEEE Press, Southampton (2006)
10. Cristianini, N., Shawe-Taylor, J.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)

Semi-supervised Learning by Spectral Mapping with Label Information

Zhong-Qiu Zhao[1], Jun Gao[1], and Xindong Wu[2]

[1] School of Computer & Information, Hefei University of Technology, Tunxi Road No.193 Hefei.Anhui 230009, China
[2] Department of Computer Science, University of Vermont, USA
zhongqiuzhao@gmail.com

Abstract. A novel version of spectral mapping for partially labeled sample classification is proposed in this paper. This new method adds the label information into the mapping process, and adopts the geodesic distance rather than Euclidean distance as the measure of the difference between two data points. The experimental results show that the proposed method yields significant benefits for partially labeled classification with respect to the previous methods.

Keywords: Semi-supervised Learning; Spectral Mapping; Geodesic Distance.

1 Introduction

In many applications of pattern classification and data mining, collecting labeled samples can be a costly and time-consuming task, whereas unlabeled ones are far easier to obtain. Therefore, semi-supervised learning has attracted much interest from researchers [1]. Recently, there have been several transductive learning algorithms based on the nearest neighbor graph on the set of high-dimensional 'inputs', such as the methods with local and global consistency [2], using Gaussian fields and harmonic functions [3], and so on. These nonlinear schemes on manifold structures can discover the intrinsic geometry of the data, and thereby propagate the label information from labeled samples to unlabeled ones.

In addition, Belkin & Niyogi [4] and Du et al. [5] presented another manifold method, referred to as spectral mapping, for partially labeled classification by developing the idea of Laplacian eigenmaps [6]. Like other schemes on manifold structures, this method assumes that nearest neighbors (according to some similarity measure) in the high-dimensional input space have similar 'outputs', thereby projecting all samples into a low-dimensional manifold. This method exploits the intrinsic structure of the data to improve classification with unlabeled examples under the assumption that the data resides on a low-dimensional manifold within a high-dimensional representation space, and this assumption seems to be reasonable in some cases. For example, a handwritten digit '0' can be fairly accurately represented as an ellipse, which is completely determined by the coordinates of its foci and the sum of the distances from the foci to any point. Thus the space of ellipses is a five-dimensional

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 448–454, 2010.

manifold. On the other hand the dimensionality of the ambient representation space is the number of pixels which is typically far higher.

Unfortunately, the label information, which can make the projection more beneficial to classification, is omitted from the mapping process. Though labeled samples may be less among the sample set, the label information, as the prior knowledge, is very important to improve the mapping efficiency, and has been utilized in the designs of energy functions of other manifold methods [2][3]. In this paper, we propose a new version of spectral mapping that makes use of the label information for semi-supervised learning. Moreover, in order to better preserve the intrinsic geometry of the data when performing spectral mapping, we adopt the geodesic distance rather than the straight-line Euclidean distance as the measure of the difference between two data points.

2 Basic Framework

Compared with the unsupervised version, our semi-supervised spectral mapping has two advantages. One is to consider for mapping not only the distances between the pairs of data points, but also their label information, though the labeled samples may be less; the basic idea for this change is to forcibly make two data points with the same class label have the strongest connection to each other among all point pairs, while two points with different class labels have the weakest connection to each other. Shown in Figure 1 is the plot of the connection between each data point pair.

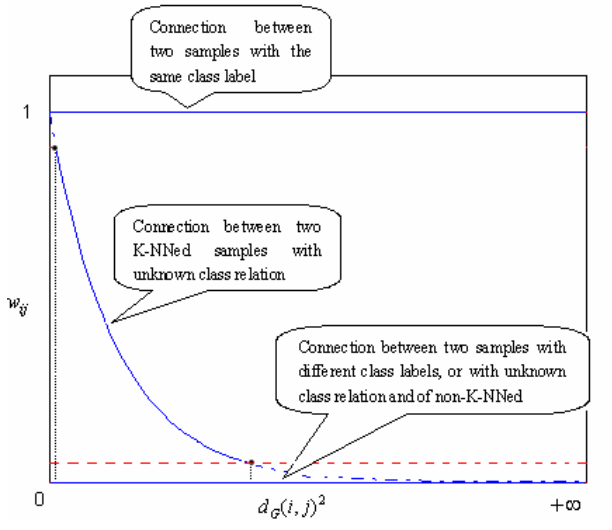


Fig. 1. The plot of the connection between each data point pair

The other advantage of our method is to use the geodesic distance rather than the straight-line Euclidean distance as the measure of the difference between two data points. Generally, on an underlying manifold, it was found that the geodesic distance along the surface of the manifold better preserves the intrinsic geometry of the data than the Euclidean distance [7]. Our proposed algorithm is summarized as follows.

Given a dataset $X=\{X_1, X_2, ..., X_l, X_{l+1}, ..., X_n\}$, the first l samples are labeled by $\{t_1, t_2, ..., t_l\}$, while the remaining samples do not have labels.

Step 1) Formation of an adjacent graph. Firstly set the parameter K. Then define the graph G with all n data points, and construct an edge to link point X_i to point X_j, with the length of

$$d(i,j)=\begin{cases}\left|X_i-X_j\right|, & if \quad X_i \overset{K}{\sim} X_j;\\ +\infty \quad , & otherwise\end{cases} \tag{1}$$

where $X_i \overset{K}{\sim} X_j$ denotes that the points X_i and X_j are among the *K*-nearest neighbors of each other.

Step 2) Estimate of the geodesic distances. (In case of a neighboring point pair, the geodesic distance is approximated by the Euclidean distance. While in case of a faraway point pair, the geodesic distance is approximated by adding up a sequence of "short hops" between neighboring points.) Firstly initialize $d_G(i,j)=d(i,j)$; for each value of $k=1,2,...,n$, in turn replace all entries $d_G(i,j)$ by $\min\{d_G(i,j), d_G(i,k)+d_G(k,j)\}$.

Step 3) Define the affinity matrix $W=\left(w_{ij}\right)$ by the following rule:

$$if \quad i\le l \ \&\ j\le l, \qquad w_{ij}=\begin{cases}1, & t_i=t_j\\ 0, & t_i\ne t_j\end{cases},$$

$$else \qquad w_{ij}=\begin{cases}\exp\left(-d_G(i,j)^2 \Big/ \beta\right), & X_i \overset{K}{\sim} X_j\\ 0, & otherwise\end{cases} \tag{2}$$

Then construct a diagonal matrix $S=\left(s_{ij}\right)$ with the same size as W, where $s_{ii}=\sum_{j=1}^{n} w_{ij}$.

Step 4) Solve the optimization problem as

$$\min_{\mathrm{y}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-y_j)^2\right\}, \quad subj.\ to \quad \sum_{i=1}^{n} s_{ii}y_i^2=1 \tag{3}$$

where $\mathrm{y}=\left(y_1, y_2, ..., y_n\right)^T$ denotes the coordinates of the first dimension for the n samples in the low-dimensional space. The solution is given by computing the top eigenvector for the eigenvector problem:

$$S^{-1}W\mathrm{y} = \lambda \mathrm{y} \tag{4}$$

Then the coordinates of all n points on a mapped m-dimensional manifold are found by computing the top $m+1$ eigenvectors of the graph Laplacian $S^{-1}W$. The top eigenvector is generally discarded since it consists of all equal components and gives no information on data structure [8].

Finally we classify an unlabeled sample by the simple nearest neighbor rule in the m-dimensional space, i.e., assign it to the class to which the nearest labeled sample belongs. Note that though two free parameters, namely K and β, appear in our algorithm, the parameter β can be easily defined. Since β is used to prevent w_{ij} from falling too fast when $d_G(i, j)$ is relatively large, its value is set to be the average geodesic distance between all pairs of data points.

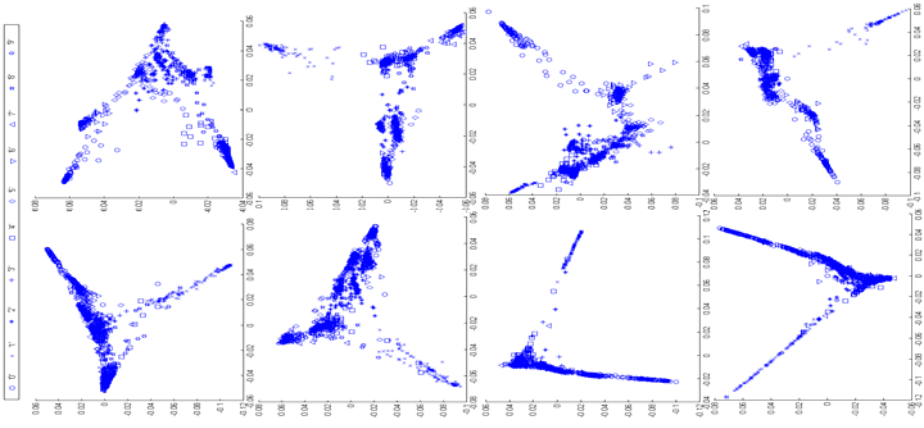


Fig. 2. Visualization of the USPS data mapped into low-dimensional spaces described by the first two coordinates. Bottom: unsupervised spectral mapping; Top: supervised spectral mapping. The 1st column: K=10, l=100; the 2nd column: K=10, l=300; the 3rd column: K=20, l=100; the 4th column: K=20, l=300.

3 Experimental Results

Our semi-supervised spectral mapping has been compared with the unsupervised version and other manifold methods on the famous USPS (U.S. Postal Service) database. The database was collected from actual handwritten postal codes, which contains 9298 normalized grey scale images of size 16*16. These images are all from 10 classes, which represent digits from 0 to 9.

In the database, all images are provided with class labels. For each trial, we randomly draw a set of 1000 samples from the database, among which a random subset is used with labels as the labeled set, and the remainder is used as the unlabeled

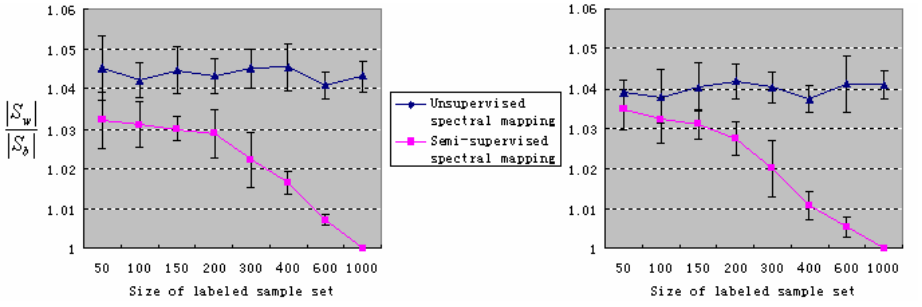


Fig. 3. Comparison of mapping performance scored by $|S_w|/|S_b|$ between unsupervised spectral mapping and semi-supervised version. Left curve chart: K=10; right curve chart: K=20. Each $|S_w|/|S_b|$ is calculated in the mapped low-dimensional space. The results are all averaged over 20 different random draws for the dataset, and the short lines on the curves show the standard deviations for the estimated values of $|S_w|/|S_b|$.

set. For all trials, we set the parameter K as $K=10$ and $K=20$, respectively. For both of the methods of spectral mapping, namely the unsupervised spectral mapping and our semi-supervised version, we set the dimensionality of the low-dimensional space, namely, m as $l/5$, which seems to work well in a series of tests.

Shown in Figure 2 is a comparison between the unsupervised spectral mapping and our semi-supervised version, of the visualizations for the 1000 data points mapped into the low-dimensional spaces described by the first two coordinates. As shown in Figure 2, our method gathers the data points from the same classes closer, and maps the points from different classes farther. To make a clearer observation, we also compute the ratio of the within-class scatter and the between-class scatter, namely $|S_w|/|S_b|$ in the mapped space, to evaluate the mapping performance. Let the within-class scatter matrix be defined as:

$$S_W = \sum_{i=1}^{c} \sum_{Y_k \in C_i} (Y_k - \overline{Y}^i)(Y_k - \overline{Y}^i)^T \tag{5}$$

and the between-class scatter matrix be defined as:

$$S_b = \sum_{i=1}^{c} n^i (\overline{Y}^i - \overline{Y})(\overline{Y}^i - \overline{Y})^T \tag{6}$$

where Y_k denotes the mapped coordinates of the sample X_k; $\overline{Y} = (1/n)\sum_{j=1}^{n} Y_j$ is the mean mapped coordinates of all samples; $\overline{Y}^i = (1/n^i)\sum_{j=1}^{n^i} Y_j^i$ is the mean mapped coordinates of the *i*th class; n^i is the number of samples in the *i*th class; and c is the number of classes. Then the comparison of the mapping performance scored

by $|S_w|/|S_b|$ between unsupervised spectral mapping and our semi-supervised version is depicted in Figure 3. These results in Figure 3 show that in the mapped spaces, our semi-supervised spectral mapping always has smaller values of $|S_w|/|S_b|$ than the unsupervised version, and with the increase of the number of labeled samples, the values of $|S_w|/|S_b|$ for our method gradually fall. Further, in Figure 4, we also summarize the average classification error rates of various semi-supervised learning methods, and the best *k*-NN (*k*=1; 3; 5) in the original space as well, for different numbers of labeled samples. It can be seen that our semi-supervised mapping method outperforms others, especially when the size of the labeled set is relatively small.

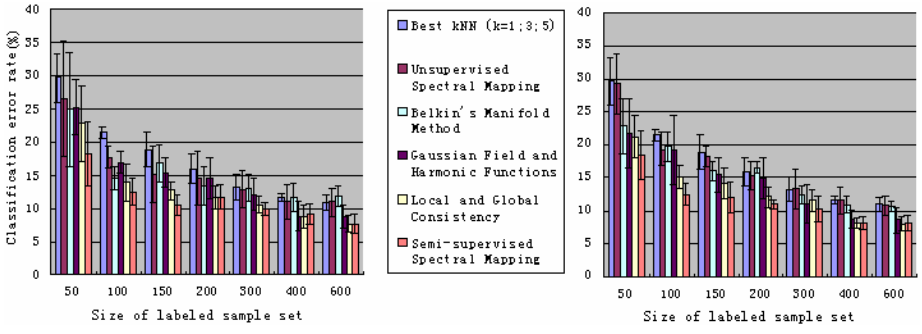


Fig. 4. Classification error rates of various semi-supervised learning methods for different numbers of labeled samples. Left bar chart: K=10; right bar chart: K=20. Each error is calculated on the unlabeled part of the dataset. The results are all averaged over 20 different random draws for the dataset, and the short lines on the bars show the standard deviations for the estimated classification error rates.

4 Conclusions

In this paper, a novel version of spectral mapping for partially labeled classification has been proposed. Our method adds the label information into the spectral mapping, and adopts the geodesic distance rather than Euclidean distance as the measure of the difference between two data points. Thereby, our new spectral mapping obtains a smaller value of $|S_w|/|S_b|$ than the original version, which implies that the data points from the same class are closerly projected to the low-dimensional manifold, while the ones form different classes are more apart projected to the low-dimensional manifold. The experimental results have shown that our approach not only improves the performance of spectral mapping, but also outperforms other semi-supervised learning algorithms.

Acknowledgements

This work has been supported by the Research Fund for the Doctoral Program of Higher Education of China (200803591024 and 20090111110015), China Postdoctoral

Science Foundation (20100470825), the National Natural Science Foundation of China (NSFC 60875012, 61005007 and 60828005), the US National Science Foundation (NSF CCF-0905337), and the 973 Program of China (2009CB326203).

References

1. Seeger, M.: Learning with labeled and unlabeled data. Technical report, The University of Edinburgh (2000)
2. Zhou, D.Y., Bousquet, O., Lal, T.N., Weston, J., Scholkopf, B.: Learning with local and global consistency, Max Planck Institute for Biological Cybernetics Technical Report (2003)
3. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian field and harmonic functions. In: Proceeding of the 20th International Conference on Machine Learning (ICML 2003), Washington DC (2003)
4. Belkin, M., Niyogi, P.: Semi-supervised learning on Riemannian manifolds. Machine Learning 56, 209–239 (2004)
5. Du, W., Inoue, K., Urahama, K.: Dimensionality reduction for semi-supervised face recognition. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 1–10. Springer, Heidelberg (2005)
6. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
7. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
8. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge Univ. Press, Cambridge (1990)

Drop Fingerprint Recognition Based on Self-Organizing Feature Map

Jie Li, Qing Song, Yuan Luo, and Cunwei Zou

Automation School,
Beijing University of Posts and Telecommunications,
Beijing 100876, China
songqing512@126.com

Abstract. Drop analysis technology developed rapidly, the recognition of drop fingerprint become more and more important. It discussed about drop analysis technology and the methods to recognize liquid drop fingerprint. With the self-learning, self-organizing and out-supervision, self-organizing feature map network is suitable to use in drop fingerprint recognition. By MATLAB simulation, a SOM neural network which has been trained is established. Two groups of samples are identified. The identification ratio of one group is 97.5 percent, and the other group is 95 percent. The recognition performance achieved the goal as expected.

Keywords: We would like to encourage you to list your keywords in this section.

1 Introduction

In recent years, Fiber-Capacitive Drop analysis technology developed rapidly, drop fingerprint identification methods are more and more, for example:

Graphics-based fingerprint recognition, the principle for this method is comparing the drop fingerprint graphics between the control group and the test group, if the difference between the graphs within the threshold, the result returns true, otherwise false.

Eigenvector-based fingerprint recognition, the principle for this method is comparing the eigenvectors between the control group and the test group. When the Euclidean distance between the eigenvectors beyond a certain value, the result returns false.

Eigenvalue-based fingerprint recognition, the principle for this method is comparing the eigenvalues between the control group and the test group.

The methods described before can only test the liquid is true or false, they can't realize the pattern recognition in drop analysis. And the eigenvalues and eigenvectors based on a large number of test data. It will reduce the efficiency of liquid recognition.

As the relationship between drop fingerprint and liquid lack of theoretical results, the color, composition and other physical and chemical characteristics is not easy to

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 455–460, 2010.

judge by the drop fingerprint. But researches show there are many difference in drop fingerprints between different liquids. The self-organizing feature map network is out-supervision, self-organizing, self-learning network. In this paper, it introduces a drop fingerprint recognition method base on self-organizing feature map network. By this method, it can realize liquid pattern recognition.

2 Drop Fingerprint

Drop Analysis Technology (DAT) is a technology that under a certain test system, use various means to implement monitoring of drop growing, in order to obtain the physical and chemical parameters for liquids and recognize the liquids from the parameters.

With DAT, the drop fingerprint can be obtained. It reveals the variation law when optical signal through the droplet with the droplet growth. Drop fingerprint reflects the measured liquid variety of physical and chemical properties, and under certain conditions, the result is exclusive.

In this paper, the drop fingerprint means Fiber-Capacitive Drop Fingerprint. The fiber and capacitive signals obtained by Fiber-Capacitive Drop Sensor, after filtering, normalization, forming drop fingerprint graphs. Fanta, Coca Cola, Sprite, Xingmu are four kinds of drinks, their drop fingerprint graphs are different (as Fig.1).

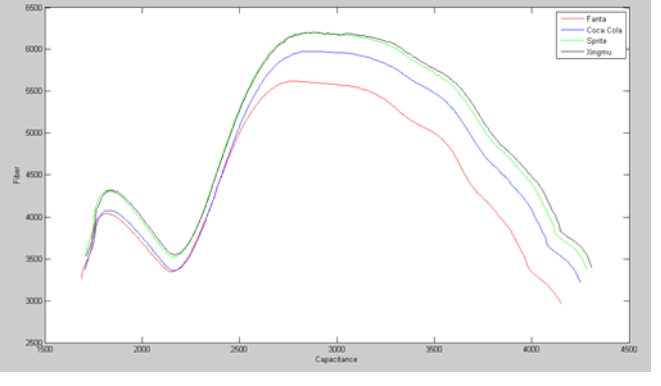


Fig. 1. Drop Fingerprint graphs of Fanta, Coca Cola, Sprite and Xingmu

3 Self-Organizing Feature Map

The Self-Organizing Feature Map (SOM) was proposed by Finnish scholar Teuvo Kohonen in 1981. This network is based on the physiology and brain science research. It is based on the biological function of nerve cells as following:

1, The sensitive cells are stimulated by external signals, by self-learning they reflect some of the physical characteristics of external stimuli, and form neurons which sensitive to some physical characteristics.

2, Under the stimulation by external signals, the cells form a kind of feature column by self-learning and self-organizing. A feature column executes a certain function.

3.1 The Structure of SOM Network

The structure include four parts:

(1) Array processing unit. Accept the event input, and form discriminant function to these signals.

(2) The mechanism of comparison and selection. Compare the discriminant function and select the processing unit with max output.

(3) Local Interconnection. Stimulant the selected processing unit and its neighboring processing unit at the same time.

(4) Adaptive process. Correct the parameters of the stimulant processing unit, in order to increase the output of discriminant function with a certain input.

The basic structure of SOM has two layers, competitive layer and input layer. The data inputted to the input layer, after weighted between the competitive layer and input layer, it returns a set of output. (as Fig.2)

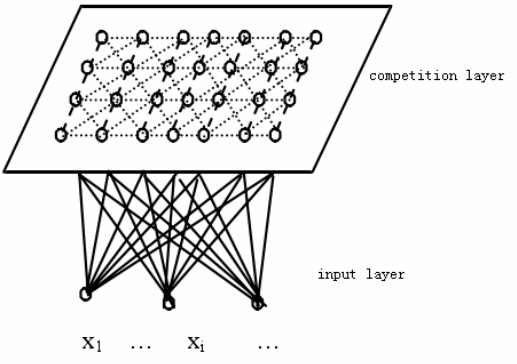


Fig. 2. Structure Model of SOM Network

3.2 Operational Summary of the SOM Algorithm

(1), Initialize the weights $W_{ij}(t)$ to some small random numbers, the value of the neighbourhood around node j $N_j(0)$, and the learning rate η_0.

(2), Pick a input vector $X(t)=[x_0(t), x_1(t), \ldots\ldots, x_{n-1}(t)]$, $x_i(t)$ is the input of node i at time t.

(3), Calculate the distance of input and output, $y_j = \sum_{i=0}^{n-1}[x_i(t)-w_{ij}(t)]^2$.

(4), Select the output node i with minimum y_j.

(5), Update η_k and $N_j(k)$.

(6), Return to (2) until there is no observable change in the map.

4 Methods

Because of the color, transparency, viscosity, surface tension and other physical characteristics of different liquids are different, their drop fingerprints are different. A drop fingerprint graph can represent a kind of liquid.

In this paper, 1000 fiber signal data in a drop fingerprint are used as an input vector X_i in SOM network. According to the demand, output nodes are set as $M \times N$ neurons. With the SOM algorithm, every input vector gets a winning neuron Y_i as output after training in the network. The output Y_i represents the drop fingerprint. Different sample vectors are training in the neural network in order to get their outputs $Y_1, Y_2, Y_3 \ldots\ldots Y_1, Y_2, Y_3$ represent different liquids. When recognize an unknown liquid, use DAT to get its drop fingerprint, and simulate in the network which has been already trained. The output of the network is Y_t. If Y_t equals to Y_i, the unknown liquid is the liquid which Y_i represent.

With the help of MATLAB Neural Network Toolbox, the code of the SOM algorithm is reduced greatly. Fig.3 and Fig.4 are the main procedure flow chart in this experiment.

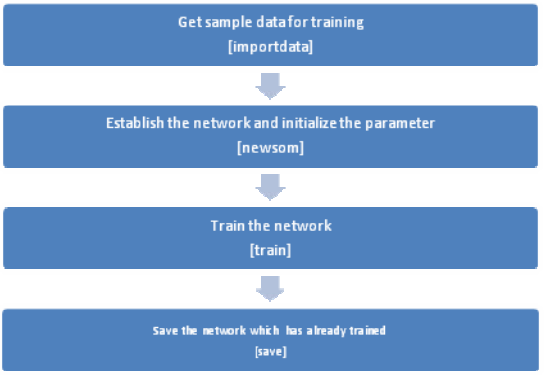


Fig. 3. The flow chart of self-organizing feature map neural network training

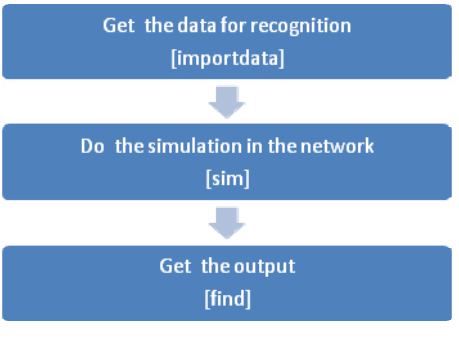


Fig. 4. The flow chart of pattern recognition

5 Results

In this paper, 40 kinds of liquid samples are training. The function to establish SOM network like this:

net=newsom(P,[30 30],'hextop','linkdist',200,1)

Taking into account the dispersion of 40 samples, the number of the neurons in output layer is 30x30. When the training is finish, the neural network realized the exact classification of 40 liquids.(as Fig.5)

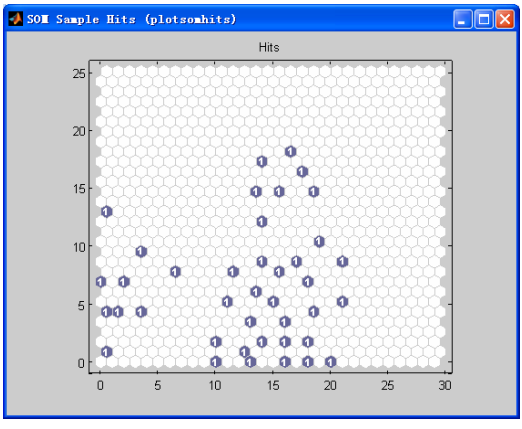


Fig. 5. Liquid classification diagram

To verify the identification ratio, the experiment is divided into two groups. One group selects 40 kinds of liquids, every liquid select a sample, 40 samples in total. In this group, Nongfu Spring is the only one which is not identified. The identification ratio is 97.5% (as Table.1). The other group selects 6 kinds of liquids, every liquid select 10 samples, 60 samples in total. In this group, 3 samples can not be identified (T on behalf of the correct identification, F on behalf of the wrong identification). The identification ratio is 95% (as Table.2). The identification ratio of this neural network is more than 95%.

Table 1. Recognition results of some samples

sample	result	sample	result	sample	result	sample	result
Nestle mineral water	Nestle mineral water	**Nongfu spring**	Unknown	**Glutinous rice wine**	Glutinous rice wine	**Five Star Beer**	Five Star Beer
Wahaha mineral water	**Wahaha mineral water**	**Shanxi mature vinegar**	Shanxi mature vinegar	**Quanxing Liquor**	Quanxing Liquor	**Yanjing Beer**	Yanjing Beer
Wahaha Purified Water	Wahaha Purified Water	**Coco Cola**	**Coco Cola**	**Erguotou liquor**	**Erguotou liquor**	**White vinegar**	White vinegar

Table 2. Recognition results of the samples

sample	1	2	3	4	5	6	7	8	9	10
Nongfu spring	T	**F**	**F**	**F**	T	T	T	T	T	T
Wahaha mineral water	T	T	T	T	T	T	T	T	T	T
Shanxi mature vinegar	T	T	T	T	T	T	T	T	T	T
Quanxing Liquor	T	T	T	T	T	T	T	T	T	T
Dark soy	T	T	T	T	T	T	T	T	T	T
Sprite	T	T	T	T	T	T	T	T	T	T

6 Conclusions

Self-organizing feature map neural network has features as automatic classification and recognition. Therefore, the successful applications in data classification are many. In this paper, SOM network is successful used in drop fingerprint recognition, and very good results obtained. It verifies the feasibility of neural network used in liquid recognition, and provides a good technology to drop fingerprint recognition. However, the neural network did not achieve 100% drop fingerprint recognition. In future research, can improve on the parameters and algorithm to obtain a more favorable neural network.

Acknowledgment

Especial thanks to The Natural Science Foundation of China (NSFC） for the financial support (Grant Number: 60702004).

References

1. Zhang, D.: MATLAB Neural Network Simulation and Application, pp. 198–204. Electronic Industry Press, Bingjing (2009)
2. Gao, J.: Theory and simulation of artificial neural networks, pp. 168–173. Machinery Industry Press (2007)
3. Song, Q., Qiu, Z., Zhang, G.: Optoelectronic drop analysis method used for liquid property study. Journal of Optoelectronics and Advanced Materials 6(4), 1213–1223 (2004)
4. Zhong, L., Rao, B., Zou, C.: Artificial Neural Network and Its Application Technology Integration, pp. 46–48. Science Press, Beijing (2007)
5. Kumar, S.: NEURAL NETWORKS, pp. 552–558. Tsinghua University Press, Beijing (2006)

Nonlinear Complementarity Problem and Solution Methods

Longquan Yong

Department of Mathematics, Shaanxi University of Technology,
Hanzhong 723001, Shaanxi, P.R. China
yonglongquan@sohu.com

Abstract. This paper provides a survey to some of recent developments in the field of nonlinear complementarity problems (NCP). Some existence conditions of solution to the NCP are given according to the monotonicity of the functions, and corresponding NCP examples are demonstrated respectively. Meanwhile, a couple of different solution methods for NCP are described. Finally, we provide a brief summary of current research trends.

Keywords: nonlinear complementarity problems, monotone function, NCP-function, Newton method, projection-type method, interior point method.

1 Introduction

The nonlinear complementarity problem (NCP), which is to find a vector $x \in \mathbb{R}^n$ satisfying the system of equations and inequalities

$$x \geq 0, \quad F(x) \geq 0, \quad x^{\mathrm{T}} F(x) = 0$$

or, equivalently,

$$x_i \geq 0, \quad F_i(x) \geq 0, \quad x_i{}^{\mathrm{T}} F_i(x) = 0 \ \text{ for } i = 1, \ldots, n.$$

Here $F : \mathrm{X} \to \mathbb{R}^n$ is a given function defined on a subset $\mathrm{X} \in \mathbb{R}^n$ containing at least the nonnegative orthant.

NCP was introduced by Cottle in his Ph.D. thesis in 1964, and the closely related Variational Inequality Problem (VIP) was introduced by Hartman and Stampacchia in 1966, primarily with the goal of computing stationary points for nonlinear programs [1-2].

Well over a thousand articles and several books have been published on the subject of complementarity problems. Many of the basic theoretical results in complementarity problems have been known for a long time. An excellent survey of much of the research in this area prior to 1990 can be found in [Harker and pang, 1990]. Further references and more recent work can also be found in [Ferris and Pang, 1997, Harker, 1993, Pang, 1994]. Here, we give just a flavor of the known existence and uniqueness results and concentrate more on the algorithmic point of view. For a general treatment of NCP and VIP we recommend Ref. [3-6]. Additional references

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 461–469, 2010.

are given in [2], as well as in the references mentioned above. We limit the scope of this paper.

In this paper, we give a survey of complementarity problems that covers some standard results and presents a number of new developments in the field. The organization of this paper is as follows. In section 2 we then state a few existence conditions for the standard NCP, and section 3 describes a couple of different solution techniques for NCP. Discussions some of applications and trends in NCP are describe in Section 4.

The notation used in this paper is assumed to be rather standard: Then n-dimensional Euclidean space is denoted by $\mathbb{R}^n$ with $\|x\|$ being the Euclidean norm of a vector $x \in \mathbb{R}^n$, and $\mathbb{R}^2_+$ denotes the set of n dimension nonnegative vectors. The scalar product of two vectors $x, y \in \mathbb{R}^n$ is written as $x^{\mathrm{T}} y$.

2 Theoretical Results for NCP

In this section, we deal with the follow questions: Does the NCP have at least one solution? When is the solution unique? Consider the following simple examples [7].

Example 2.1. In all examples gives here, the function $F : \mathbb{R}^1 \to \mathbb{R}^1$ is defined.

(a) If $F(x) = -x - 1$, then NCP has no solution since $F(x) < 0$ for all $x \geq 0$.
(b) If $F(x) = x - 1$, then NCP has a unique solution $x^* = 1$.
(c) If $F(x) \equiv 0$, then NCP has infinitely many solutions $x^* \in [0, +\infty)$.
(d) If $F(x) = (x-1)^2 - 1$, then NCP has two isolated solutions $x^* = 0$ and $x^* = 2$.

In order to state some existence and, possibly, uniqueness results, we now introduce some classes of functions that play the role of corresponding convexity properties in optimization problems.

Definition 2.2. Let $\mathrm{X} \subseteq \mathbb{R}^n$ be a nonempty set and $F : \mathrm{X} \to \mathbb{R}^n$ be a given function. Then F is said to be

(i) monotone if $(x - y)^{\mathrm{T}} (F(x) - F(y)) \geq 0$ for all $x, y \in \mathrm{X}$;
(ii) strictly monotone if $(x-y)^{\mathrm{T}}(F(x)-F(y)) > 0$ for all $x, y \in \mathrm{X}$ with $x \neq y$;
(iii) strongly monotone if there exists a constant $\mu > 0$ (called modulus) such that $(x - y)^{\mathrm{T}} (F(x) - F(y)) \geq \mu \|x - y\|^2$ for all $x, y \in \mathrm{X}$.
Using these definitions, we can state the following result.

Theorem 2.3. Let $F : \mathrm{X} \to \mathbb{R}^n$ be a continuous function on $\mathrm{X} = \mathbb{R}^n_+$. Then the following statements hold:

(a) If F is monotone on X, then NCP has a convex (possibly empty) solution set.
(b) If F is strictly monotone on X, then NCP has at most one solution.
(c) If F is strongly monotone on X, then NCP has a unique solution.

Definition 2.4. A vector $x \in \mathbb{R}^n$ is called feasible for NCP if $x \geq 0$ and $F(x) \geq 0$, and a vector $\hat{x} \in \mathbb{R}^n$ is called strictly feasible for NCP if $\hat{x} > 0$ and $F(\hat{x}) > 0$.

Theorem 2.5. Let $F : \mathrm{X} \to \mathbb{R}^n$ be a continuous function on $\mathrm{X} = \mathbb{R}^n_+$. Assume that F is monotone on X and there is a strictly feasible point $\hat{x}$ for NCP. Then NCP has a nonempty and compact solution set.

Following we give some examples of NCP.

Example 2.2. [8]**Example of unique solution.**

Consider the NCP, where $F : \mathbb{R}^5 \to \mathbb{R}^5$ is given by

$$F(x) = \begin{pmatrix} x_1 + x_2x_3x_4x_5/50 \\ x_2 + x_1x_3x_4x_5/50 - 3 \\ x_3 + x_1x_2x_4x_5/50_4 - 1 \\ x_4 + x_1x_2x_3x_5/50 + 1/2 \\ x_5 + x_1x_2x_3x_4/50 \end{pmatrix}$$

The NCP has only one solution $x^* = (0,3,1,0,0)^T$.

Example 2.3. [9, Kojima–Shindo Poblem] **Example of two solutions.**

Consider the NCP, where $F : \mathbb{R}^4 \to \mathbb{R}^4$ is given by

$$F(x) = \begin{pmatrix} 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 - 6 \\ 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2 \\ 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 3x_4 - 9 \\ x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3 \end{pmatrix}$$

This is a degenerate NCP and has two solutions $x^* = (\sqrt{6}/2,0,0,0.5)^T$ and $x^* = (1,0,3,0)^T$.

Example 2.4. [10-11] **Example of alternative solutions.**

Consider the NCP, where $F : \mathbb{R}^4 \to \mathbb{R}^4$ is given by

$$F(x) = \begin{pmatrix} -x_2 + x_3 + x_4 \\ x_1 - (4.5x_3 + 2.7x_4)/(x_2 + 1) \\ 5 - x_1 - (0.5x_3 + 0.3x_4)/(x_3 + 1) \\ 3 - x_1 \end{pmatrix}$$

This problem has infinitely many solutions $x^* = (\lambda, 0, 0, 0)^T$, where $\lambda \in [0,3]$.

Remark 2.1

If $F : \mathbb{R}^2_+ \to \mathbb{R}^2$ is monotone, and there exists feasible point (but not exists strictly feasible), then NCP maybe unsolvable.

For example, let $F(x)=\left(2x_1x_1-2x_2+1, x_1^2+2x_1-1\right)^T$. Since Jacobian matrix $F'(x)$ is positive semi-definite in $\mathbb{R}^2_+$, $F:\mathbb{R}^2_+\to\mathbb{R}^2$ is monotone. The feasible region $S=\left\{x\middle|x\geq 0, x\geq 0\right\}=\left\{x\in\mathbb{R}^2_+\middle|x_1=1, x_2\geq 0\right\}$. For any feasible point $x\in S$ we have $x^TF(x)=1$, thus this monotone NCP is unsolvable.

Remark 2.2

If $F:\mathbb{R}^2_+\to\mathbb{R}^2$ is strictly monotone and the feasible point of NCP exists, then NCP maybe unsolvable.

For example, let $F(x)=\left(\varphi(x_1)+x_2, x_2\right)^T$, where $\varphi:\mathbb{R}^1\to\mathbb{R}^1$ is monotone increasing and for any $t\in\mathbb{R}^1$, $\varphi(t)<0$. It is easy prove that $F:\mathbb{R}^2_+\to\mathbb{R}^2$ is strictly monotone and feasible point exists. If any feasible point $x'=(x_1', x_2')\geq 0$ satisfies $(x')^TF(x')=0$, then $x_1'=x_2'=0$. However, $F(0)=(\varphi(0),0)^T\notin\mathbb{R}^2_+$, thus this monotone NCP is unsolvable.

3 Algorithms for NCP

3.1 Projection-Type Methods

Projection-Type Methods are based on the following simple observation.

Proposition 3.1. *A* vector x^* is a solution of NCP if and only if x^* satisfies the fixed-point equation

$$x=(x-\gamma F(x))_+,$$

where $\gamma>0$ is an arbitrary constant and z_+ denotes the Euclidean projection of a vector $z\in IR^n$ onto the nonnegative orthant.

Projection-Type methods typically attempt a fixed-point iteration. The simplest consists of the iteration

$$x^{k+1}:=(x^k-\gamma F((x^k-\gamma F(x^k))_+))_+,\quad k=0,1,2,\cdots,$$

Where $x^0\in\mathbb{R}^n_+$ is any given starting point. An excellent survey of much of projection-type methods can be found in [12].

3.2 Merit Functions Methods

We first give the definition of a merit function.

Definition 3.2. *A* function $\Psi:\mathbb{R}^n\to\mathbb{R}$ is called a merit function for NCP if it has the following properties:

(a) $\Psi(x) \geq 0$ for all $x \in \mathbb{R}^n$;

(b) $\Psi(x) = 0$ if and only if x is a solution of NCP.

This definition suggests replacing the NCP by the unconstrained minimization problem

$$\min \psi(x), \quad x \in \mathbb{R}^n.$$

More and new merit function can be found in [13].

3.3 Nonsmooth Newton Methods

It is possible to reformulate the NCP as a nonlinear system of equations using so-called NCP-functions.

Definition 3.3. A function $\phi : \mathbb{R}^2 \to \mathbb{R}$ is called an NCP-function if it satisfies the condition $\phi(a,b) = 0 \Leftrightarrow a \geq 0, b \geq 0, ab = 0$.

The following functions are NCP-functions:

(a) $\phi(a,b) = \min\{a,b\}$;

(b) $\phi(a,b) = \sqrt{a^2 + b^2} - a - b$;

(c) $\phi(a,b) = \lambda\left(\sqrt{a^2+b^2} - a - b\right) - (1-\lambda)\max\{0,a\}\max\{0,b\}$ with $\lambda \in (0,1)$ fixed;

(d) $\phi(a,b) = -ab + \frac{1}{2}\min^2\{0, a+b\}$.

Normally (a) is called the minimum function, and (b) is called the Fischer-Burmeister function. Example (c) is termed the penalized Fischer-Burmeister function. Example (d) is the only NCP-function that is differentiable on the whole space. Many other NCP-functions are known in the literature [14-15].

For any given NCP-function ϕ, define $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Phi(x) := \left(\phi(x_1, F_1(x)), \cdots, \phi(x_n, F_n(x))\right)^T.$$

If F and the NCP-function ϕ are continuously differentiable, then the same holds for the equation operator Φ. In this case, we can apply the classical Newton method to $\Phi(x) = 0$ to solve the underlying NCP [16-17]. This leads to the iteration

$$x^{k+1} := x^k - \Phi'\left(x^k\right)^{-1} \Phi\left(x^k\right), k = 0,1,2\cdots$$

for a given starting point $x^0 \in \mathbb{R}^n$. We can expect local fast convergence if the standard assumptions for the Newton method are satisfied; in particular, the Jacobian

matrix $\Phi'(x^*)$ has to be nonsingular at a solution x^* of NCP. Unfortunately, Kanzow shows that this matrix is singular at any degenerate solution. One way to overcome the singularity problem is to use smooth NCP-function.

3.4 Smooth Newton Methods

Smoothing Newton method for nonlinear complementarity problem has become a research focus. J.Y HAN, L.Q.Qi, X.Q.Yang, N.H.XIU, J-S Chen, Z.H.Huang, etc. has done a lot in this pioneering work. Smoothing methods for the NCP typically approximate a nonsmooth NCP-function ϕ by a smooth function ϕ_μ that depends on a smoothing parameter $\mu > 0$. For example, it is possible to approximate the Fischer-Burmeister function by

$$\phi(a,b) = \sqrt{a^2+b^2} - a - b \approx \sqrt{a^2+b^2+\mu} - a - b =: \phi_\mu(a,b).$$

Using similar techniques, it is possible to smooth most known NCP-functions.

We introduce a class of smoothing functions as

$$\phi_\theta(\mu,a,b) = \sqrt{\theta(a-b)^2 + (1-\theta)(a^2+b^2) + 4\mu} - a - b, \theta \in [0,1].$$

Obviously, Fischer-Burmeister smoothing function and CHKS smoothing function are special cases when $\theta = 0$ and $\theta = 1$ respectively.

As soon as we have such a smooth approximation ϕ_μ, we define the corresponding equation operator

$$\Phi_\mu(x) := \left(\phi_\mu(x_1, F_1(x)), \cdots, \phi_\mu(x_n, F_n(x))\right)^T$$

which is now a smooth approximation of Φ if F itself is smooth.

The main idea of any smoothing method is to apply a standard Newton method to the nonlinear system of equations $\Phi_\mu(x) = 0$ and to drive the smoothing parameter μ down to zero. If we denote the value of μ at the *k*th iteration by μ_k, a typical iteration of a smoothing method is given by

$$x^{k+1} = x^k - [\Phi'_{\mu_k}(x^k)]^{-1}\Phi_{\mu_k}(x^k), \quad k = 0,1,2,\cdots$$

A critical point of all smoothing methods is the way the smoothing parameter μ_k gets updated at each iteration. If this is done in an appropriate way, it is possible to prove global convergence for monotone problems, and even local fast convergence if a solution x^* of the NCP satisfies the strong regularity assumption. More detail smoothing Newton method can be found in [18-19].

3.5 Interior Point Methods

Another important class of methods used to tackle NCP is interior point methods.

So far, the theory developed for interior point methods is much stronger than the one developed for smoothing methods. Interior point methods have a common feature

such that generates a sequence $\{x^k, k = 0,1,2,\cdots\}$ in the feasible region under the assumption of knowing a feasible initial point x^0. If each point x^k of the generated sequence satisfies the equality system $F(x^k) \geq 0$, then we say that the algorithm is a feasible interior point algorithm.

However, it is a very difficult task to find a feasible initial point to start the algorithm. To overcome this difficulty, recent studies have been focused on some new interior point algorithms without the need to find a feasible initial point. In 1993, Kojima et al. presented the first infeasible interior point algorithm with global convergence for linear programming, soon later Zhang and Wright introduced this technique to the linear complementarity problem. Interior-point methods can often be shown to have polynomial complexity and local fast convergence even in the absence of the strong regularity condition [20-22]. An excellent survey of much of the research in this area can be found in [23].

3.6 Linearization Methods

Here, the fundamental idea is to linearize F about the current iterate x^k and solve the corresponding linear complementarity problem (LCP)

$$x \geq 0, \;\; F(x^k) + F'(x^k)\left(x - x^k\right) \geq 0, \;\; x^T\left(F(x^k) + F'(x^k)(x - x^k)\right) = 0$$

to generate x^{k+1}. In the context of generalized equations, this method, which has been commonly termed SLCP (for sequential LCP), was shown to be locally well-defined and fast convergent under the strong regularity condition [7].

3.7 Multiobjective Optimization Methods

In [24], general linear complementarity problems are studied via multiobjective optimization methods. For NCP we can similarly formulate the NCP into a multiobjective optimization problem (MOP) and calculate the zero-efficient solution of the MOP, which is also the solution to the NCP. An excellent survey of much of the research in this area can be found in [25-26].

4 Applications and Current Trends of NCP

NCP arise naturally in the study of many phenomena in economics and engineering. While excellent algorithms have been developed, there is still much attention being devoted to developing new algorithms for NCP. Mixed complementarity problems, parameter complementarity problems, robust complementarity problems, variational inequalities and mathematical programs with equilibrium constraints are current research focuses.

Acknowledgments. The author is very grateful to the referees for their valuable comments and suggestions. This work is supported by Natural Science Foundation of Shaanxi Educational Committee (No.09JK381).

References

1. Lemke, C.E., Howson, J.T.: Equilibrium points of bimatrix games. SIAM J. Appl. Math. 12, 413–423 (1964)
2. Billups, S.C., Murty, K.G.: Complementarity Problems. Journal of Computational and Applied Mathematics 124, 303–318 (2000)
3. Ferris, M.C., Pang, J.S. (eds.): Complementarity and Variational Problems: State of the Art. SIAM Publications, Philadelphia (1997)
4. Ferris, M.C., Pang, J.S.: Engineering and economic applications of complementarity problems. SIAM Review 39, 669–713 (1997)
5. Ferris, M.C., Pang, J.S., Ferris, M.C., Pang, J.S. (eds.): Complementarity and Variational Problems: State of the Art. SIAM Publications, Philadelphia (1997)
6. Harker, P.T., Pang, J.S., Harker, P.T., Pang, J.S.: Finite-dimensional variational inequality and nonlinear complementarity problems, a survey review of theory, algorithms and applications. Math. Prog. 48, 161–220 (1990)
7. Ferris, M.C., Kanzow, C.: Complementarity and related problems: A survey. In: Pardalos, P.M., Resende, M.G.C. (eds.) Handbook of Applied Optimization, pp. 514–530. Oxford University Press, New York (2002)
8. Xia, Y., Leung, H., Wang, J.: A projection neural network and its application to constrained optimization problems. IEEE Transactions on Circuits and Systems–I 49, 447–458 (2002)
9. Kojima, M., Shindo, S.: Extensions of Newton and quasi-Newton methods to systems of PC equations. Journal of Operations Research Society of Japan 29, 352–374 (1986)
10. Jiang, H., Qi, L.: A New Nonsmooth Equations Approach to Nonlinear Complementarity Problems. SIAM J. Control Optim. 35, 178–193 (1997)
11. Kanzowa, C.: Some equation-based methods for the nonlinear complementarity problem. Optimization Methods and Software 3, 327–340 (1994)
12. Xiu, N.H., Zhang, J.: Some recent advances in projection-type methods for variational inequalities. J. Comput. Appl. Math. 152, 559–571 (2003)
13. Du, S., Gao, Y.: Merit functions for nonsmooth complementarity problems and related descent algorithm. Applied Mathematics - A Journal of Chinese Universities 25, 78–84 (2010)
14. Sun, D., Qi, L.: On NCP-Functions. Computational Optimization and Applications 13, 201–220 (1999)
15. Chen, J.-S., Pan, S.: A family of NCP-functions and a descent method for the nonlinear complementarity problem. Computational Optimization and Applications 40, 389–404 (2008)
16. Li, D., Fukushima, M.: Smoothing Newton and Quasi-Newton Methods for Mixed Complementarity Problems. Comput. Optim. Appl. 17, 203–230 (2000)
17. Lopes, V.L.R., Martínez, J.M., Pérez, R., Lopes, R.: On the Local Convergence of Quasi-Newton Methods for Nonlinear Complementarity Problems. Appl. Numer Math. 30, 3–7 (1999)
18. Qi, L., Sun, D.: A survey of some nonsmooth equations and smoothing Newton methods. In: Eberhard, A., Glover, B., Hill, R., Ralph, D. (eds.) Progress in optimization, pp. 121–146 (1999)
19. Qi, L., Sun, D., Zhou, G.: A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities. Mathematical Programming 87, 1–35 (2000)

20. Kojima, M., Mizuno, S., Yoshise, A.: A Polynomial-time Algorithm for a Class of Linear Complementarity Problems. Math. Prog. 44, 1–26 (1989)
21. Zhang, Y.: On the Convergence of a Class of Infeasible-Interior-Point Methods for the Horizontal Linear Complementarity Problem. SIMA J. Optim. 4, 208–227 (1994)
22. Wright, S.J.: An Infeasible-Interior-Point Algorithm for Linear Complementarily Problems. Math. Prog. 67, 29–52 (1994)
23. Potra, F.A., Wright, S.J.: Interior-point methods. J. Comput. Appl. Math. 124, 255–281 (2000)
24. Isac, G., Kostreva, M.M., Wiecek, M.M.: Multiple objective approximation of feasible but unsolvable linear complementarity problems. Journal of Optimization Theory and Applications 86, 389–405 (1995)
25. Kostreva, M.M., Wiecek, M.M.: Linear complementarity problems and multiple objective programming. Mathematical Programming 60, 349–359 (1993)
26. Kostreva, M.M., Yang, X.Q.: Unified Approaches for Solvable and Unsolvable. Linear Complementarity Problems, European Journal of Operational Research 158, 409–417 (2004)

A New Path Planning Method for a Shape-Shifting Robot

Mengxin Li[1], Ying Zhang[1], TongLin Liu[2], and Chengdong Wu[2]

[1] School of Information and Control Engineering, Shenyang Jianzhu University , Shenyang 110168, China
limengxin@sjzu.edu.cn, zy_1120@sjzu.edu.cn
[2] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
tlliu@sin.cn, wuchengdong@ise.neu.edu.cn

Abstract. A shape-shifting robot with changeable configurations can accomplish search and rescue tasks which could not be achieved by manpower sometimes. The accessibility of this robot to uneven environment was efficiently enlarged by changing its configuration. In this paper, a path planning method is presented that integrates the reconfigurable ability of the robot with the potential field law. An adaptive genetic algorithm is applied to solve effectively the local minimum problem. The experiments show that the robot's configurations can be changed to perform the path planning with the environmental variation. Moreover, the path has been shortened effectively.

Keywords: Reconfiguration, Shape-shifting, Path planning, Robot.

1 Introduction

To achieve different tasks and adapt various environments such as search and rescue, military reconnaissance, and mine exploration, the flexibility of the traditional robots is too worse to meet the needs. A reconfigurable mobile robot has attracted a lot of attention and interest for wide applications. Because of automatic reconfigurable ability, it can complete the search and rescue task by changing the configurations to adapt with the dangerous, unknown environment. Up to now, there have been a lot of research groups outside China which study reconfigurable mobile robot: USC-ISI, CMU, AIST , Fukuda Laboratory and et al. In China, scientists in Shanghai Jiao Tong University Shenyang Institute of Automation, Chinese Academy of Sciences, Beijing university of aeronautics and astronautics, Harbin Institute of Technology, and et al. [1,2].

The path planning of robot is key research topics, which determines a robot whether can perform completing tasks on time in complex environment. At present, many path planning algorithms have been researched, such as Rolling Path Planning Method, Genetic Algorithm, Ant Colony Algorithm and Visibility Graph [3]. One of the well known methods is the potential field method with rapid description of the environment [4,5], which is widely used for real-time obstacle avoidance and motion planning. It minimizes a penalty function that consists of two components describing influence of the obstacles in the workspace and the intention to go to the desired point.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 470–478, 2010.

2 System of the Shape-Shifting Robot

The shape-shifting robot with the automatic reconfigurable ability adopts modular structure. The adaptability of the robot is reinforced in the different space by changing the configurations.

The six ultrasonic sensors are emplaced in three regions of the platform, 2 sensors are fixed on the left and right of frontage respectively, 2 sensors are fixed on the front and rear of right side respectively, and 2 sensors are fixed on the front and rear of left side respectively. The detecting-distance of the every ultrasonic sensor is 4cm-5m, and they are asynchronously transmitted. An electron compass, a GPS and an inclinometer are emplaced in the platform respectively. The shape-shifting robot has diverse configurations. The reconfigurable capacity of the robot will be reinforced by using every motion joint. Four kinds of commonly used configuration of the shape-shifting robot, i.e. configuration R, T, d, L.

The control system is composed of a wireless module, a main control module, motor control modules and sensor-based feedback control modules.

Main control module is the robot's decision-making unit, which estimates its own position, and plans path and changes its own configuration according to information of environment provided by the feedback control module.

The sensor-based feedback system contains controllers 8 to 11 as well as electron compass, GPS, inclinometer and ultrasonic sensors. The controller 11 receives the data of every controllers of ultrasonic sensors by CAN bus and processes them. All sensors receive information from the environment and transmit the data to respective controllers in which pre-settled algorithms are used to deal with them, and then results are sent to the PC.

3 Modified Potential Field Method

3.1 Calculation of Attractive and Repulsive Forces

In the motion space, the robot is considered as one particle, and the particle receives effect of potential field force, according to the basic idea of potential field method. In the potential field approach every obstacle exerts a repulsive force on the robot and target position exerts an attractive force. According to characteristic which the shape-shifting robot has reconfigurable capacity, the shape-shifting robot is considered as a robot which will be reduced proportionally, and keep the reconfigurable ability which change the configurations to strengthen the accessibility.

The amplitude of the repulsive force is defined as:

$$\boldsymbol{F}_r = w_1 e^{-c_1(d_r^2+a)} \tag{1}$$

Where w_1, c_1 and a are constant coefficients respectively. d_r is distance between the current robot location and the obstacle location. Its value is calculated as follows:

$$d_r = \sqrt{\Delta x_o^2 + \Delta y_o^2} = \sqrt{(x-x_o)^2 + (y-y_o)^2} \tag{2}$$

where $\triangle x_o$ and $\triangle y_o$ are coordinate distances between the robot position and the obstacle position.

We proposed the following law for attractive force:

$$\boldsymbol{F}_a = \begin{cases} w_2 d_a & 6 < | X - X_t | \\ k_1 w_2 d_a & 3 < | X - X_t | \leq 6 \\ k_2 w_2 d_a & | X - X_t | \leq 3 \end{cases} \tag{3}$$

where X is current position coordinate of the robot. X_t is position coordinate of target. w_2, k_1 and k_2 are constant coefficients respectively. The k_1 and k_2 are involved to reach the target when the distance d_a between the current robot location and the target location are located into the circle with radius (3, 6] m or 3m. The value d_a is calculated as follows:

$$d_a = \sqrt{\Delta x_t^2 + \Delta y_t^2} = \sqrt{(x - x_t)^2 + (y - y_t)^2} \tag{4}$$

The force controlled on the robot：

$$\boldsymbol{F}_s = min(\boldsymbol{F}_r + \boldsymbol{F}_a) \tag{5}$$

where the resulting vector $\boldsymbol{F}_s$ is obtained as a minimum of the vector sum of repulsive force $\boldsymbol{F}_r$ and attractive force $\boldsymbol{F}_a$.

3.2 Method of Accessibility

In terms of complicated environments and irregular polygonal obstacles, if stacking potential fields of all obstacles, the time consumption for computation would be too large to perform the task in real-time application [6]. The concept of the visibility field is introduced according to location of sensors. The shape-shifting robot only needs to avoid obstacles detected.

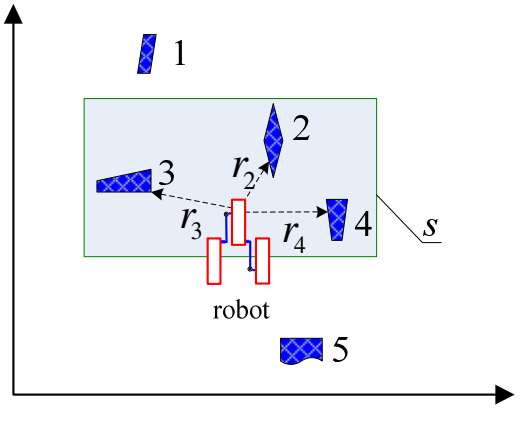


Fig. 1. The visibility field of the robot. s is the scope of the visibility field, 1, 2, 3, 4 and 5 is obstacles.

The robot's scope of the visibility field here is a rectangle with the length a=8m and width b=4m according to measuring range of ultrasonic sensor, as shown in Figure 1, where obstacles 2, 3 and 4 are visible, obstacles 1 and 5 is not visible.

In order to further reduce the computation load, Firstly, nearest obstacle from the robot, the distance of the obstacles and obstacles which locate along both sides of robot's path are all considered to satisfy the width of a certain configuration of the robot. If there is only one obstacle in the scope of the visibility field, it is very easy for avoidance. If there are many obstacles in the scope of the visibility filed, the robot needs to calculate the maximum resultant forces $\boldsymbol{F}_{r\max}$ of two nearest repulsive forces before going through the two obstacles.

$$\boldsymbol{F}_{r\max} = \boldsymbol{F}_{r1} + \boldsymbol{F}_{r2} \tag{6}$$

where $\boldsymbol{F}_{r1}$ and $\boldsymbol{F}_{r2}$ are repulsive forces of two nearest obstacles which the robot intends to pass through.

The current control force on the robot:

$$\boldsymbol{F}_s = min(\boldsymbol{F}_{r\max} + \boldsymbol{F}_a) \tag{7}$$

Corollary 1. Due to the effect of resultant repulsive force $\boldsymbol{F}_{r\max}$ when control force of (7) is implemented, we can derive the rule that the robot movements towards the centric position of two obstacles and goes through along the path near the centric position.

According to (1), (6) and corollary 1, we can get the relationship between configuration of the robot and maximum repulsive force $\boldsymbol{F}_{r\max}$, as represented in Table 1. s_T, s_d and s_L are halves of width of configuration T, d and L respectively. Δd_1 and Δd_2 are smaller adjustment quantity. Thus the robot can automatically select a proper configuration that satisfies the distance between two obstacles.

In Figure 2(a), the shape-shifting robot doesn't adopt reconfigurable method. The robot discovers that obstacle 2 is the nearest from robot, when it moves to the target. The robot hopes that obstacle 2 is bypassed. The path of first choice is l_1, but the robot discovers that the distance between obstacle 2 and obstacle 5 is less than the width of robot with the configuration T. The path l_2 of secondary choice is in the same way. So the robot had chosen the farther path l_3. In Figure 2(b), the shape-shifting robot with two kinds of reconfigurations can change its own configuration to choose the nearer

Table 1. Condition of the robot changing the configuration

configuration	condition
T	$\boldsymbol{F}_{r\max} < w_1 e^{-c_1((s_T+\Delta d_1)^2+a)} + w_1 e^{-c_1((s_T+\Delta d_2)^2+a)}$
d	$\boldsymbol{F}_{r\max} \geq w_1 e^{-c_1((s_T+\Delta d_1)^2+a)} + w_1 e^{-c_1((s_T+\Delta d_2)^2+a)}$
	$\boldsymbol{F}_{r\max} < w_1 e^{-c_1((s_d+\Delta d_1)^2+a)} + w_1 e^{-c_1((s_d+\Delta d_2)^2+a)}$
L	$\boldsymbol{F}_{r\max} \geq w_1 e^{-c_1((s_d+\Delta d_1)^2+a)} + w_1 e^{-c_1((s_d+\Delta d_2)^2+a)}$
	$\boldsymbol{F}_{r\max} < w_1 e^{-c_1((s_L+\Delta d_1)^2+a)} + w_1 e^{-c_1((s_L+\Delta d_2)^2+a)}$

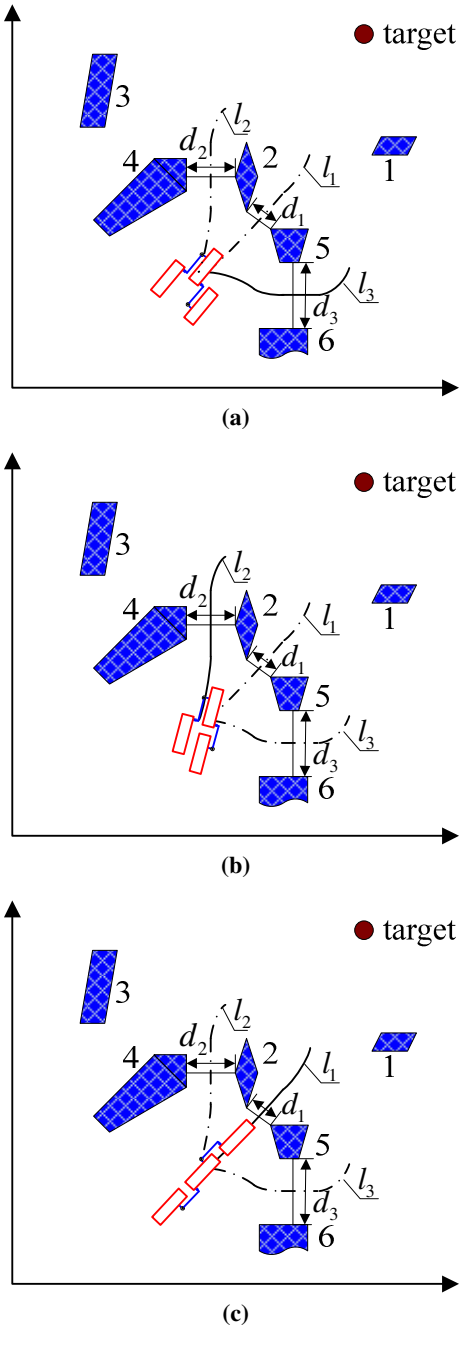


Fig. 2. Robot's passing through the narrow space

path l_2 and can pass through obstacles successfully according to the environmental variation. In Figure 2(c), the shape-shifting robot with three kinds of reconfigurations can change its own configuration to choose the nearest path l_1. Therefore, the better the reconfigurable ability of the robot is, the better the adaptive ability of environment is.

3.3 Solving the Local Minimum Problem

In some case, the robot only knows the surrounding environment of current location in complex environment, so it falls into the local minimum easily. Therefore this paper adopts an adaptive genetic algorithm (AGA) as following to estimate and escape from the local minimum:

AGA can choose crossover operator and mutation operator adaptively according to fitness, which result in morerapidly convergence speed comparing with traditional GA. It is more important AGA is used to avoid local minimum.

Crossover operator, Pc, and mutation operator, Pm have influence on behavior of GA, which are changed with fitness. When all the fitnesses of population are tended to be consistent. Pc and Pm are added. When all the fitnesses are dispersed, Pc and Pm are reduced.. The individuals which have lower Pc and Pm corresponding to higher fitness than average fitness are protected to go to next generation. While the individuals which have relatively highly Pc and Pm corresponding to lower fitness than average fitness are removed.

Pc and Pm are given by：

$$P_c = \begin{cases} P_{c1} - \dfrac{(P_{c1} - P_{c2})(f' - f_{av})}{f_{\max} - f_{av}} & f' \geq f_{av} \\ P_{c1} & f' < f_{av} \end{cases} \tag{8}$$

$$P_m = \begin{cases} P_{m1} - \dfrac{(P_{m1} - P_{m2})(f_{\max} - f)}{f_{\max} - f_{av}} & f' \geq f_{av} \\ P_{m1} & f' < f_{av} \end{cases} \tag{9}$$

where f is fitness, $f'$is derivative of f,; fav is average of f, $fmax$ is max of f, and P_{c1}, P_{c2} P_{m1} P_{m2} are fitted to 0.9,0.6,0.1,0.001 respectively.

4 Experiment of Accessibility and Simulation of Path Planning

4.1 Experiment of Accessibility

In order to validate effectiveness of the method, we do the experimental research on the autonomic accessibility of the shape-shifting robot. In an unknown condition of obstacles, the robot can change its movement route for obstacles avoidance according to the information of sensors. Also, the robot can estimate the distance between the neighboring obstacles. Then, it automatically adjusts its own configuration to reinforce the adaptability of environment.

4.2 Simulation of Path Planning

The simulation experiments have been done in the same complex environment with polygonal obstacles using the method proposed, as represented in Figure 3, Figure 4 and Figure 5. In these figures, the red rectangle is start position, the red circle is target position.

In Figure 3, with the modified potential field method, the robot with the fixed configuration T completed path planning efficiently in the environment with polygonal obstacles.

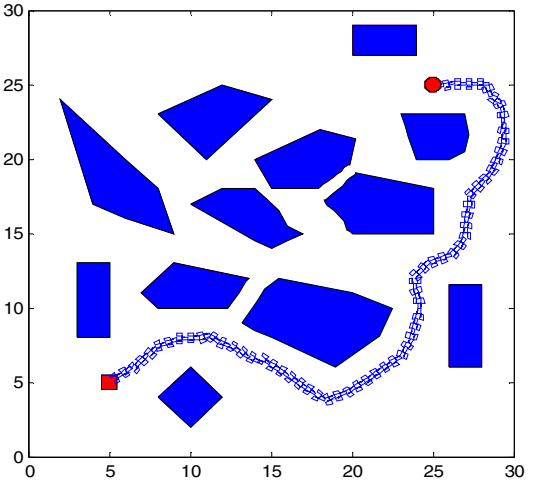


Fig. 3. Robot path planning with the configuration T

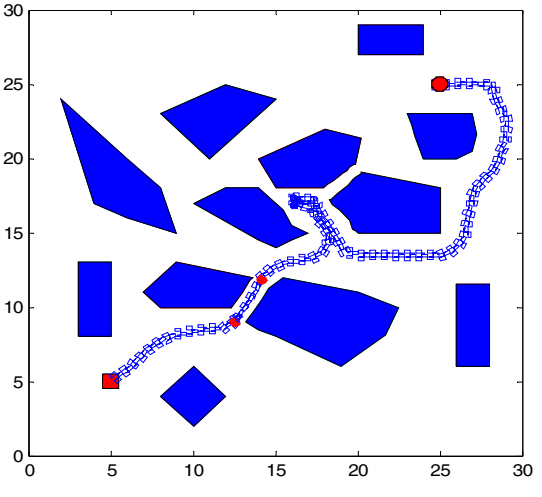


Fig. 4. Robot path planning with two kinds of reconfigurations according to environmental variation

In Figure 4, the robot with two kinds of reconfigurations can change its own configuration to complete path planning according to the environmental variation, after integrating the reconfigurable ability of the robot into the modified algorithm. We find out: the method well solved the local minimum problem in irregular obstacles.

In Figure 5, the robot with three kinds of configurations can change its own configuration to perform path planning more effectively according to the environmental variation than before. When the robot passed through first narrow space, it transformed its own configuration from T to d. While the robot passed through secondary narrower space, it changed its own configuration from T to d to L.

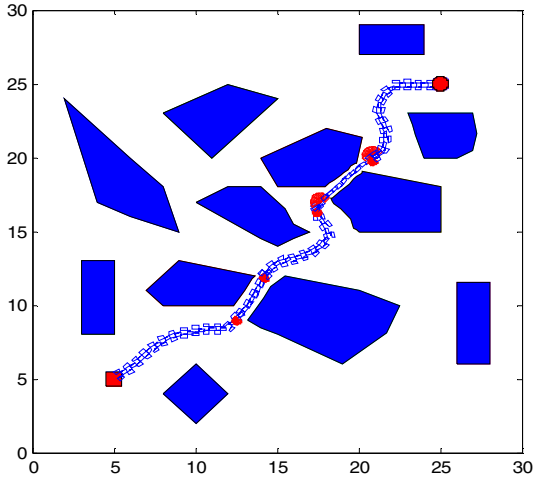


Fig. 5. Robot path planning with three kinds of reconfigurations according to environmental variation

5 Discussion and Conclusions

In this paper, a modified potential field method is presented to describe environment rapidly. Moreover, with AGA, the local minimum problem is solved to make sure real-time application effectively. In addition, the proposed method has a very strong capability of acclimatization. It has been validated through experiment results that the robot can change its own configuration to perform the path planning with the environmental variation. Furthermore the method proposed can shorten the path effectively.

References

1. Yim, M., Shen, W., Malemi, B., et al.: Modular Self-Reconfigurable Robot Systems. IEEE Robotics & Automation Magazine, 43–52 (2007)
2. Bhat, P.S., Kuffner, J., Goldstein, S., et al.: Hierarchical Motion Planning for Self-Reconfigurable Modular Robots. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp. 886–891 (2006)

3. Kamimura, A., Kurokawa, H., Yoshida, E., et al.: Automatic Locomotion Design and Experiments for A Modular Robotic System. IEEE-ASME Trans Mechatron, 314–325 (2005)
4. Takagawa, I., Sekiyama, K., Fukuda, T.: Coevolution of Physical Configuration and Control Strategy on Self-Reconfigurable Flexible Transfer System. In: Proceedings of IEEE International Conference on Intelligent Robot and Systems, pp. 2474–2479 (2003)
5. Wu, Q.X., Cao, G.Y., Fei, Y.Q.: Motion Simulation And Experiment of A Novel Modular Self-Reconfigurable Robot. J. Southeast Univ., 185–190 (2006)
6. Fei, Y.Q., Zhao, X.F.: Modules Classification and Automatic Generation of Kinematics on Self-Reconfigurable Modular Machines. J. Intell. Robot. Syst., 147–159 (2005)

Pedestrian Gait Classification Based on Hidden Markov Models

Weihua Wang and Zhijing Liu

School of Computer Science and Technology
Xidian University
No.2 South TaiBai Road Xi'an, Shaanxi, 710071, China
wangwh2005@gmail.com,
Liuzhijing@vip.163.com

Abstract. Analysis of human activity from video sequences is one of the hottest and difficult research areas in computer visions. Because of the fact that human continuous motion can be decomposed into an image sequence based on time, state space method is applied in this paper. First, Silhouettes are extracted using the Background Subtraction method and features are represented by moment. Then a method using recursion method for establishment of the standard gait state sequence is proposed. In order to determine whether the behavior is abnormal in different scenarios, wavelet moment is used to extract features of the human body images, and then recognizes the moving human bodies activity based on Discrete Hidden Markov Model. The experiment tests show some encouraging results also indicates the algorithm has very small leak-examining and mistake-examining-rate which indicate that the method could be a choice for solving the problem but more tests are required.

Keywords: Human Motion Recognition; Hidden Markov Models; style; Gait Classification; wavelet moment.

1 Introduction

In recent years, as a sharp increase in terrorist attacks on crowded public places, like airports, bus stations, subways and markets etc, surveillance systems have been more demanded. As one of the significant objectives of such systems, human activity understanding and recognition which is one of the most active research topics in the field of computer vision, has a number of promising applications, ranging from security and surveillance to augmented reality. For this reason, it is important to develop intelligent visual surveillance to replace the traditional passive video surveillance that is proved to be ineffective as the number of cameras exceeds the capability of human operators to monitor them.

All of the visual surveillances are evidence of a great and growing interest in detecting moving walker. Because, the detection of moving walkers could help to detect crimes, accidents, terrorist attacks, but it is hard to define "abnormal" in different scenarios. The most convenient way is to classify the pedestrians' gaits in the beginning, and then behavior validity can be judged under special scenarios.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 479–487, 2010.

Gait classification attempts to use gait as biometric information to understand people by the characteristics of their motion. It also includes efforts to classify different types of human activities, for example, walk, faint, run, bench, jump, and wander action. Research in the field has raised many methods: they are model-based methods [1][2], statistical methods [3] and spatial-temporal motion-based methods [4]. And Invariant moments due to its character are a popular tool for extracting feature of moving object. The application of classical moments to two dimensional images was first shown in the early sixties by Hu [5]. In [6] Lee and Grimson computed a set of images features that are based on moments such as centroid, major axis of each part of people. Shutler et al. developed new Zernike velocity moments to describe the motion throughout an image sequence to help recognize gait [7].

Although considerable achievements in this field have been accomplished in recent years, some challenges still remain to be overcome. The optimal gait classification should classify motion with real-time. What we do in this paper is a spatio-temporal motion-based model method with little prior knowledge, and it belongs to the work of the behavior understanding. Our paper therefore offers two main contributions:

1. A method using multiple features of the moving silhouette for establishing different standard motion base is presented.
2. We use Hidden Markov Models to build human posture detection model and classify the body postures into a number of categories simply with centroid tracking.

The paper is organized as follows: Section 2 describes a method for silhouette extraction. In Section 3, using moving features extracted by wavelet moment a motion standard sequence is established and Hidden Markov Models is introduced to build human posture detection model and recognize the human motion in Section 4. Realization of the system and experimental results are presented in Section 5, followed by the conclusions part.

2 Contour Extraction

In common, human body models can be classified into 2-D models in the image plane and 3-D models in real world coordinates. Due to their simplicity, 2-D models have more applications. In this paper, classification is performed outdoors, applying a 2-D model with a single fixed camera.

2.1 Background Modeling

An algorithm based on the statistical Gaussian model is applied to estimate the background image, which is composed of initialization and update. In an interval time M , the mean and the covariance of the brightness of every pixel are computed as the initially estimated background image, that is $B_0 = [\mu_0, \sigma_0^2]$, where:

$$\mu_0(x, y) = \frac{1}{M} \sum_{i=0}^{M-1} f_i(x, y) \tag{1}$$

$$\sigma_0^2(x, y) = \frac{1}{M} \sum_{i=0}^{M-1} [f_i(x, y) - \mu_0(x, y)]^2 \tag{2}$$

After initializing the estimated background image, with the appearance of new image, the parameters of background image should be updated self-adaptively according to the following formulas. If the updated image was $B_t = [\mu_t, \sigma_t^2]$, then its updating rules are:

$$\mu_t = (1-\alpha)\mu_{t-1} + \alpha \times f_t \tag{3}$$

$$\sigma_t^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(f_t - \mu_t)^2 \tag{4}$$

Where α is a given constant and range from 0 to 1.

The frequency of background updating is depended on the rate of light variety. The variety frequency of light is following:

$$\lambda(t) = \left| 1 - \sum_{q \in Q} \frac{|B(q)||I(q)|\cos\theta}{|B(q)|} \right| \tag{5}$$

Where θ is the angle between the colour vector and background vector. $B(q)$ is the lightness of current location in background model. $I(q)$ is lightness of current location in the frame. The background model update when the threshold is larger than λ.

2.2 The Detection and Extraction of Moving Object

The core issues of moving target detection and segment is making use of motion detection algorithm to obtain change regions from background image in video sequence, and on this basis object segmentation is performed. The most popular methods of object detection are the temporal differencing and the background subtraction. The former is of good adaptability to dynamic environments, efficient and easy to implement, but it cannot extract all the related points of object integrally; whereas the latter can extract the object points rather integrally by comparing the foreground image with background image, but it is too sensitive to dynamic scene change caused by illumination and external conditions such as light and weather changes. In order to overcome the factors of these two moving object detection algorithms, we propose a weighted average method based on the temporal difference and the background subtraction, and the precise extraction of moving object contour is implanted successfully and the detection result is quite good, as shown in Fig.1.

Fig. 1. The result of human contour extraction

3 Motion Base

The process of human behavior recognition could simply be considered as the classification of time-varying data, and the essential step is the matching test motion sequence with standard motion sequence. The key issues of recognition is how to obtain standard motion sequence from learning sample sequence, and meanwhile the standard sequence is able to handle slight feature changes in spatial and temporal scale. Thus, an algorithm for image similarity based on Wavelet moments is proposed.

3.1 Wavelet Velocity Moments

The concept of moments comes from physics, where it relates to the force required to affect a given angular acceleration on an object of known mass distribution. For the purpose of image analysis, a set of generalized moments which represent different features of the target shape has been extended and combined with the theory of algebraic invariants to become the mathematical device used today in image analysis, such as HU moments and Zernike moments etc. Both of invariant moments show good performances at global information extraction using the characteristic of image translation, rotation and scale invariant. However they are not efficient in classifying similar objects into different category, although the higher order moments can reflect the differences in similar images, but it is hard to obtain.

Wavelet moments [8] are perfect representations of moments in multi-resolution wavelet domain, which integrates the theory of moment invariants into wavelet analysis and can strengthen the analysis ability to subtle image feature. Typically, Complex wavelet moments of scaling factor m and translating factor n are defined as:

$$W_{m,n,q} = \iint f(r,\theta)[\Psi_{m,n}(r)e^{jq\theta}]^{*} r dr d\theta \tag{6}$$

where * denotes the complex conjugate, $f(r,\theta)$ is the function described in polar coordinates. q is a parameter, representing feature of the image object. $\Psi_{m,n}(r)$ represents wavelet basis functions:

$$\Psi_{m,n}(r) = 2^{m/2}\Psi\left(2^{m}r - \frac{n}{2}\right) \qquad where\ m = 1.2.3..., n = 0.1...2^{m+1} \tag{7}$$

and according to its local optimum and orthogonal in space-frequency, the cubic B-spline wavelets function is chose as generating function of $\Psi_{m,n}(r)$, which is:

$$\Psi(r) = \frac{4a^{n+1}}{\sqrt{2\pi(n+1)}}\sigma_w \cos\left(2\pi f_0(2r-1)\right)\cdot\exp\left(-\frac{(2r-1)^2}{2\sigma_w^2(n+1)}\right) \tag{8}$$

Wavelet Velocity Moments [9] are essentially a sum of Wavelet moments over a sequence of T frames, weighted by a real-valued scalar function of the displacement of the centre of mass (CoM) between consecutive frames:

$$W_{m,n,q,\beta,\lambda} = \sum_{i=2}^{T} \sum_{r} \sum_{\theta} U(i,\beta,\lambda) f(r,\theta) [\Psi_{m,n}(r) e^{jq\theta}]^* \tag{9}$$

U is the series of weights derived from the displacements of the CoM, Usually (β,λ) are set to (0, 1) or (1, 0) to avoid zero weights derived from the displacement of the CoM when there is only horizontal or vertical motion of the pedestrians. Normalized Wavelet velocity moments $\bar{W}_{mnq\beta\lambda}$ defined in equation, where A is the average area of the moving object, T is the length of the video sequence.

$$\bar{W}_{mnq\beta\lambda} = \frac{W_{mnq\beta\lambda}}{A * T} \tag{10}$$

3.2 Feature Matching

In this paper, a distance metric between the video images and the standard image in database is made, from which, the Euclidean distance is obtained. If the test sequence set and train sequence set are respectively U and V, the similarity between them is:

$$d_{UV} = \frac{\sum_{i=1}^{k} (W_U(i) - W_V(i))^2}{K} \tag{11}$$

K is the number of cubic B-spline wavelet moments, and $W_U(i)$ is the Ith feature of cubic W_U of sequence U, which is normalized by divided by the maximal value of this feature to make it be in the range of [0,1].

3.3 Standard Motion Sequence

Normal gait information of person changes in space and temporal, which could be considered as periodic motion, so a set of discrete frames are used to express a cycle

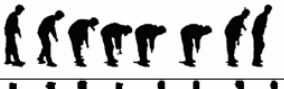
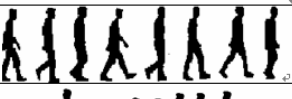
Motion	Motion Sequence
Bend	
Walk	
Jump	
Faint	
Crouch	

Fig. 2. The result of gait base

of canonical behaviors, distill people's contours, and establish libraries for them. In the stage of training, according to the result of the feature matching, the recursion method of minimal Euclidean distance is adopted to find the candidate image to establish libraries for each motion. The results of gait base are shown in Fig. 2.

4 Time Series Analysis of Human Posture

HMM is widely used to forecast the potential sequence of an event according to its samples. It is very efficient in learning and matching behavior models. There are three basic questions about HMM including evaluation, decoding and learning. The decoding means working out the most possible state sequence for the present model and observation sequence. And in this paper we define each static frame as a state, take a movement sequence as a combination of every state of different sample frames, and use combination probability as judge rule of activity belongingness to implement matching recognition of human activity.

Describe the HMM using the model which has five factors and is expressed as $\lambda = (N, M, \pi, A, B)$.The factor N represents the amount of human actions, M is the deputy of the character's total number in each action, π is the probability in which a certain action is chosen, A is the transfer probability, B represents the character distribution in each action. For specific model, in order to train HMM to recognize its observation sequence $O^i = o_1^i, o_2^i, ...o_r^i$, we can utilize Bayesian theory and Forward Algorithm to estimate the probability of observation sequence $\mathrm{P(O \mid \lambda)} = \sum_{i=1}^{n} a_T(i)$, where N is the Markov statuses number in model. Suppose forward variable is the probability of HMM being at state S_i, in observation sequence $\mathrm{o_1 o_2 \ldots o_T}$ at time t, that is $a_t(i) = \mathrm{P(o_1 o_2 \ldots o_t}, S_t = q_i \mid \lambda)$. We need to adjust parameter λ to let the appearance probability of observation $\mathrm{P(O \mid \lambda)}$ be maximal. At this stage, the number of different human postures determines the quantity of HMM coded symbols (the possible state value M). Each activity associates with a HMM, which means HMM number is never equal to the number of different activities. Hence, Maximum Likelihood Estimation Algorithm is used to estimate λ again and again in order to obtain a better λ' .Then we can figure out the maximum $P_i (i \in 1, 2, \ldots, \mathrm{M})$ among $P_1, P_2, \ldots, P_M$, which is used to recognize human gesture of current frame.

5 Evaluation and Experimental Results

For verifying the ability of proposed method, we validate the effectiveness of presented algorithm through movement analysis of walk, jump, running, crouch, wandering, and bend. The process of the method includes the following steps:

(1) In order to ensure the integrality of the contour, after we have got the person's contour by using background subtraction and temporal differencing method, also region inosculating combined with the theory of morphologic erosion and expansion is employed [10].

(2) Establish a coordinate, using circular search; find out the point which has the maximal abscissa and y-axis and the point which has the minimal abscissa and y-axis. The line connected with the two points is the rectangle's diagonal which contains the person's contour. The point of rectangle's diagonal is regard as centroid.

(3) To solve the problem when people enter or leave the scene the contour is not full, we cut off the frames about the process. Detect the motion while the target entered the scene completely.

(4) The method for classification is shown in flow chart:

If the vertical centroid is stable:

$v_x \geq 3m/s$, then the movement may be classified as running.

Else If the centroid vertical varying dramatically:

The action is jump

Else

Calculate moments feature similarity.

Use the HMM model classification algorithm.

(4) In the training stage, we assume that there is a single person performing only one action per video which length is 4800 frames, and extract relative moment features as training set for HMM model. For the experimental evaluation, HU moment and Zernike moment are also applied to make a comparison with the proposed method. The recognition results of the testing data are shown in Table 1.

Table 1. Testing Data Recognition Result

Human activities	**Recognition rate (%)**		
	HU moment	*Zernike moment*	*Method in this paper*
Walk	76%	76.7%	83%
Jump	93%	95%	94%
Crouch	82%	80%	87%
Running	70%	80%	82%
Wander	75%	80%	83%
Bend	70%	76.7%	73%
Average	77.6%	81.4%	83.6%

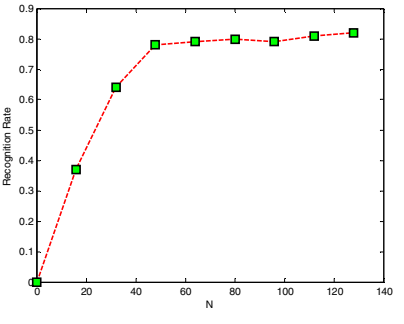


Fig. 3. The relationship between the length and the recognition rate

In the experiment, if a video with only a single gait, the correct recognition rate is high. If a video contains a number of actions, the final recognition result is the gait which has the longest duration action. In other aspects, with a single gait video the longer of video's length (N) is, the higher of recognition rate achieve. The relationship between the length and the recognition rate is following.

6 Conclusion

In this paper, we imply cubic B-spline wavelet velocity moments and HMM model to describe and recognize human gait. Comparing with some invariant moments, a Cubic B-spline wavelet velocity moment has the trait of multi-resolution analysis, and provides the feature from coarsely to finely, which considers not only the static appearance feature, but also the motion velocity features. Also experimental results have proved its superiority by showing a better recognition rate. And the good characteristics of wavelet moment will have a good prospect in the field of pattern recognition.

However, when several people enter the background at the same time, defilade will be a problem. In that case, we can screen it by several cameras from different directions, and adopt the technology of data fusion with multi-vision to obtain more information about human body. Besides, how to select better feature that contains more information and reduces computation complexity is another research task remained to be done.

Acknowledgments. Sincere thanks go to the Ministry of Education of the People's Republic of China, Teach and Research Project (Guangdong Province) for its financial support. Project No. 2006D90704017.

References

1. Wang, L., Suter, D.: Visual learning and recognition of sequential data manifolds with applications to human movement analysis. Computer Vision and Image Understanding 110, 153–172 (2008)
2. Chan, C.S., Liu, H., Brown, D.J.: Recognition of Human Motion From Qualitative Normalised Templates. Journal of Intelligent and Robotic Systems 48, 79–95 (2007)
3. Raskin, L., Rivlin, E., Rudzsky, M.: Using Gaussian Processes for Human Tracking and Action Classification. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Paragios, N., Tanveer, S.-M., Ju, T., Liu, Z., Coquillart, S., Cruz-Neira, C., Müller, T., Malzbender, T. (eds.) ISVC 2007, Part I. LNCS, vol. 4841, pp. 36–45. Springer, Heidelberg (2007)
4. Zhi-Lan, H., Fan, J., Gui-Jin, W., Xing-Gang1, L., Hong, Y.: Anomaly Detection Based on Motion Direction. ACTA AUTOMATICA SINICA 34(11) (2008)
5. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(1), 179–187 (1962)
6. Lee, L., Grimson, W.E.L.: Gait Appearance for Recognition. Biometric Authentication, 143–154 (2002)
7. Shutler, J.D., Nixon, M.S.: Zernike velocity moments for sequence-based description of moving features. Journal of Image and Vision Computing 24, 343–356 (2006)

8. Zhao, G.Y., Li, Z.B., Deng, Y.: Human motion recognition and simulation based on retrieval. Journal of Computer Research and Development 43(2), 368–374 (2006) (in Chinese with English abstract)
9. Zhao, G.Y., Cui, L., Li, H.: Combining Wavelet Velocity Moments and Reflective Symmetry for Gait Recognition. In: Li, S.Z., Sun, Z., Tan, T., Pankanti, S., Chollet, G., Zhang, D. (eds.) IWBRS 2005. LNCS, vol. 3781, pp. 205–212. Springer, Heidelberg (2005)
10. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Publishing House of Electronics Industry, Bingjing (2002)

L-Infinity Norm Minimization in the Multiview Triangulation

Yang Min

College of Automation,
Nanjing University of Posts and Telecommunications
Nanjing, China
yangm@njupt.edu.cn

Abstract. Triangulation is an important part of numerous computer vision systems. The multiview triangulation problem is often solved by minimizing a cost function that combines the reprojection errors in the 2D images. In this paper, we show how to recast multiview triangulation as quasi-convex optimization under the L-infinity norm. It is shown that the L-infinity norm cost function is significantly simpler than the L2 cost. In particular L-infinity norm minimization involves finding the minimum of a cost function with a single global minimum on a convex parameter domain. These problems can be efficiently solved using second-order cone programming. We carried out experiment with real data to show that L-infinity norm minimization provides a more accurate estimate and superior to previous approaches.

Keywords: multiview triangulation; L-infinity norm minimization; global optimal algorithms; second-order cone programming.

1 Introduction

Reconstructing the three-dimensional structure of a scene using images is a fundamental problem in computer vision. The geometric aspects of 3D reconstruction have been well-understood for a decade, but the involved optimization problems are known to be highly non-convex and difficult to solve. Traditionally, these problems are tackled using heuristic initializations followed by local, gradient-based optimization algorithms, which are prone to being enmeshed in local minima.

The L_∞ optimization is a rather new and promising direction of research in multi-view geometry [1, 2, 3]. Many important geometric problems can be formulated under the L_∞ norm by an iterative, globally optimal binary search through smaller quasi-convex minimizations of the physically meaningful geometric errors. In just a few years' time it has attracted much attention from the vision geometry community. So, this paper presents the global optimization method for estimating 3D scene structure, and this method is to replace L_2 norm with L_∞ norm. This way it is possible to obtain a provably optimal solution with a geometrically sound cost function in a relatively efficient way.

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 488–494, 2010.

2 The Triangulation Problem

Triangulation, referring to the act of reconstructing the 3D location of a point given its images in two or more known views, is a fundamental part of numerous computer vision systems. Although conceptually simple, this problem is not completely solved in the general case of n views and noisy measurements

Let $\mathbf{P}_i, i = 1, \cdots, n$ be a sequence of n known cameras, and $\mathbf{x}_i$ be the image of some unknown point $\mathbf{X}$ in 3-space, both expressed in homogeneous coordinates. Thus, we write

$$\mathbf{x}_i = \lambda_i \mathbf{P}_i \mathbf{X} \tag{1}$$

where the scalar λ_i, called the projective depth, accounts for scale. The problem of computing the point $\mathbf{X}$ given the camera matrices $\mathbf{P}_i$ and the image points $\mathbf{x}_i$ is known as the triangulation problem.

Let $\mathbf{P} = \begin{bmatrix} \mathbf{p_1} & \mathbf{p_2} & \mathbf{p_3} \end{bmatrix}^T$ denote the 3×4 camera where $\mathbf{p_i}$ are 4-vectors, $(u, v)^T$ stand for image coordinates and $\mathbf{X}$ be a homogeneous 3D point. Then the reprojection residual vector for one image is given by

$$r = \left(u - \frac{\mathbf{p_1^T X}}{\mathbf{p_3^T X}}, v - \frac{\mathbf{p_2^T X}}{\mathbf{p_3^T X}} \right)^T \tag{2}$$

Under an indepedent, identically distributed (**iid**) Gaussian noise model, the objective function to minimize is the sum-of-squared residuals, which becomes

$$\sum_{i=1}^{N} \|r_i\|^2 \tag{3}$$

where N is the number of residual terms in the problem, that is, the number of images of the given point.

Assuming independent Gaussian noise, a least-squares formulation on reprojection error and temporal consistency under the L_2 norm provides the maximum likelihood estimate which is statistically the optimal solution for this problem. However, the least-squares formulation leads to a nonlinear optimization problem with many local optima. In general, to overcome this problem, linear algorithms minimizing an algebraic error are used to initialize the nonlinear bundle adjustment.

Minimizing the sum-of-squares objective function (3) is known to be a troublesome non-convex optimization problem for both structure and transformation estimation [6]. It is known that the two-view triangulation problem can be reduced to finding the roots of a sixth degree polynomial [4], while three-view triangulation can be posed as the solution to a polynomial system which may have up to 47 roots [5]. So, not only are these seemingly simple instances of the triangulation problem known to have several local minima, the difficulty of obtaining a solution rises sharply with the number of

views. This phenomena causes difficulties for local optimization techniques (such as Newton-based methods) since they might get stuck in local minima.

3 Traditional Approaches

3.1 Linear Solution

Often, it is possible to algebraically solve for the unknown 3D point in the case of triangulation, by using a linear method. Let $(x_i, y_i)^T$ be the image of unknown point $\mathbf{X}$ under camera $\mathbf{P_i}$. Then, the following relationships hold true:

$$\begin{aligned} x_i\left(\mathbf{p}_{\mathbf{3i}}^{\mathbf{T}}\mathbf{X}\right)-\left(\mathbf{p}_{\mathbf{1i}}^{\mathbf{T}}\mathbf{X}\right)=0 \\ y_i\left(\mathbf{p}_{\mathbf{3i}}^{\mathbf{T}}\mathbf{X}\right)-\left(\mathbf{p}_{\mathbf{2i}}^{\mathbf{T}}\mathbf{X}\right)=0 \end{aligned} \tag{4}$$

If $m \geq 2$ views are available, these equations can be expressed in matrix form as

$$\mathbf{AX}=0, where\ A=\begin{bmatrix} x_1\mathbf{p}_{\mathbf{31}}^{\mathbf{T}}-\mathbf{p}_{\mathbf{11}}^{\mathbf{T}} \\ y_1\mathbf{p}_{\mathbf{31}}^{\mathbf{T}}-\mathbf{p}_{\mathbf{21}}^{\mathbf{T}} \\ \vdots \\ x_m\mathbf{p}_{\mathbf{3m}}^{\mathbf{T}}-\mathbf{p}_{\mathbf{1m}}^{\mathbf{T}} \\ y_m\mathbf{p}_{\mathbf{3m}}^{\mathbf{T}}-\mathbf{p}_{\mathbf{2m}}^{\mathbf{T}} \end{bmatrix} \tag{5}$$

and the triangulation problem can be solved by minimizing the algebraic error

$$\min_X \|\mathbf{AX}\| \tag{6}$$

Note that $\mathbf{X}=0$ is a solution to (6), so to extract a solution, we need to fix the scale of X. This can be achieved by demanding that $\|\mathbf{X}\|=1$ or by setting the last coordinate of $\mathbf{X}$ to 1. The former, which is a homogeneous version, can be solved using singular value decomposition (**SVD**), while the latter inhomogeneous version represents a linear least squares problem. The inhomogeneous version, of course, precludes points at infinity from being a solution.

Such solutions are often called Direct Linear Transform (**DLT**) methods in the literature [6]. While they are fast and applicable for a large number of views, they minimize an algebraic distance and not the geometric reprojection error in (3). So, the solution yielded by these methods may not correspond to geometric intuition and can vary depending on the choice of normalization. Moreover, the solution quality can degrade dramatically in the presence of noise. In practice, they are used as initialization to a nonlinear minimization routine, called bundle adjustment.

3.2 Bundle Adjustment

Research in Structure from Motion resulted in the emergence of a dominant accepted technique – bundle adjustment[7] . Bundle adjustment refers to a class of local iterative optimization approaches which can be used to minimize the cost function in (3). While bundle adjustment is a local optimization approach, it is quite powerful due to its flexibility and incorporation of features specific to multiview geometry. The most notable bundle adjustment methods in computer vision employ variants of the Levenberg-Marquardt iterative algorithm.

In the case of triangulation, the cost has the form (3) where the minimization is only over the structure variables. Bundle adjustment minimizes a Maximum Likelihood criterion and has the advantage of being very flexible in incorporating different kinds of variables and constraints. The disadvantage, of course, is that it is very prone to getting stuck in local minima, so it requires a strong initialization to produce meaningful estimates.

4 The L_∞ Minimization

The key idea of the L_∞ scheme is to replace the L_2 error norm with the L_∞-norm (i.e. minimax norm). Previous works show that this leads to quasi-convex minimization. We call a function quasi-convex if all of its sublevel sets are convex. Quasi-convex minimization in multi-view geometry often takes the following form ([2]):

$$\min_x \max_i f_i(x) = \frac{\left\|\mathbf{A_i x} + \mathbf{b_i}\right\|}{\left(\mathbf{c_i^T x} + \mathbf{d_i}\right)} \; s.t. \mathbf{c_i^T x} + \mathbf{d_i} > 0, i = 1, \cdots, n \tag{7}$$

where the $f_i(x)$ are quasi-convex, $x \in \mathbf{R^n}$ is the unknowns to be solved for. The dimension of the problem is n, which is often fixed and intrinsic to particular application. For example, n = 3 for multi-view triangulation. The SOCP(Second-Order Cone Programming) problem is easily solvable using commonly available software packages. We use the SeDuMi software, implemented in C with a Matlab interface [9].

The standard approach to solve such a quasi-convex program is to convert it to iteratively solving the following SOCPs via bisection:

$$\begin{aligned} &\min_x \gamma \\ &s.t. C_i(x) = \left\|\mathbf{A_i x} + \mathbf{b_i}\right\| - \gamma\left(\mathbf{c_i^T x} + \mathbf{d_i}\right) \le 0, i = 1, \cdots, n \end{aligned} \tag{8}$$

where $C_i(x)$ represents the i-th second-order cone. So (3) becomes

$$\begin{aligned} &\left\| \begin{matrix} u\mathbf{p_3^T X} - \mathbf{p_1^T X} \\ v\mathbf{p_3^T X} - \mathbf{p_2^T X} \end{matrix} \right\| = \sqrt{(u\mathbf{p_3^T X} - \mathbf{p_1^T X})^2 + \left(v\mathbf{p_3^T X} - \mathbf{p_2^T X}\right)} \\ &= \|r\| \left|\mathbf{p_3^T X}\right| \end{aligned} \tag{9}$$

We can avoid the absolute value on the right hand side of Eq. (9), since the cheirality condition $\mathbf{p}_3^{\mathrm{T}}\mathbf{X} > 0$ holds for all X. This comes from the fact that the points must physically lie in front of the cameras observing them. Then, one can rewrite Eq. (9) to

$$\left\|\mathbf{A}_\mathbf{i}\mathbf{x} + \mathbf{b}_\mathbf{i}\right\| = \gamma_i\left(\mathbf{c}_\mathbf{i}^{\mathrm{T}}\mathbf{x} + \mathbf{d}_\mathbf{i}\right) \tag{10}$$

with known $\mathbf{A}_\mathbf{i}, \mathbf{b}_\mathbf{i}, \mathbf{c}_\mathbf{i}$ and $\mathbf{d}_\mathbf{i}$, and a vector x of unknown variables and unknown γ_i.

The reprojection error for one point in Eq. (3) is a non-linear but quasi-convex function [8]. However, when summing squares of reprojections errors for all projected points, i.e. many quasi-convex functions, as done in the standard bundle adjustment, one gets a multi-modal residual function with many local minima. To prevent such a sum, and keep the problem quasi-convex with the guaranty of a global minimum,the use of L_∞ norm was proposed. The reprojection error in Eq. (3) under the L_∞ norm is defined as $\gamma = \max\left\|r_i\right\|^2$, γ_i and the problem of searching for unknown variables reads as

$$x^* = \arg\min_x \gamma = \arg\min_i \max_i \gamma_i \tag{11}$$

5 Experiments

We carried out experiment with real data to show performance of the proposed SOCP solution. For solving the SOCP we use the publicly available toolbox SeDuMi [9]. For comparison, we also applied standard linear algorithms and bundle adjustment (which optimizes the L_2-norm) to exactly the same data.

In order to test the triangulation method, the camera matrices need to be precomputed. For easier comparison, the Root Mean Squares (RMS) errors of the reprojected and measured points are given instead of the sum of squares errors. In experiment we measured RMS reprojection and L_∞ errors.

Fig. 1 shows three test images (of sizes 512 × 512) in a corridor and 4 pairs of manually selected point correspondences. And the camera matrices available at http://www.robots.ox.ac.uk/~vgg/data.html.

Fig. 1. Three test images and 4 pairs of manually selected point correspondences

Table 1 shows experimental results of two methods. After solving these problems, the RMS errors for the traditional approache and L_∞ minimization are 0.40495 pixels and 0.38783 pixels, respectively. The corresponding L_∞ errors are 0.89991 pixels and 0.85414 pixels.

Table 1. Experimental results of two methods

	Traditional Approache	L_∞ Minimization
View 1 RMS	0.44343	0.43906
View 2 RMS	0.35416	0.38123
View 3 RMS	0.41216	0.33632
Total RMS	0.40495	0.38783
L-infinity error	0.89991	0.85414

Experimental results show that the L_∞ minimization method gives better results than the traditional approache method.

6 Conclusion

L_2 cost functions usually encountered in multiview triangulation have intrinsically non-linear characteristics. These cost functions are also typically highly non-convex, so the search space is riddled with local minima. In this paper, the L_∞ norm minimization can be used with success to estimate 3D scene structure. The approach discussed here uses Second-Order Cone Programming to obtain a global solution. A critical virtue of the L_∞ scheme is that the solution obtained is not only geometrically meaningful, but also globally optimal and hence unique. This forms a sharp contrast to the conventional L_2 method, which is known to be problematic due to local minima or slow convergency.

The method introduced into the L_∞ computation provides a marked improvement, and would appear to be necessary for the best results. We show how to recast multiview triangulation problem as quasi-convex optimization under the L_∞ solvable with Second Order Cone Programming. Our formulation allows a global optimal solution to be found with a clear physical meaning of the cost function being minimized.

References

[1] Hartley, R., Schaffalitzky, F.: L_∞ minimization in geometric reconstruction problems. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 504–509 (2004)

[2] Kahl, F., Kahl, R.H.F.: Multiple view geometry under the L_∞-norm. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(9), 1603–1617 (2008)

[3] Ke, Q., Kanade, T.: Quasiconvex optimization for robust geometric reconstruction. In: Proc. 10th Int'l Conf. Computer Vision, pp. 986–993 (2005)

[4] Hartley, R.I., Sturm, P.: Triangulation. Computer Vision and Image Understanding 68(2), 146–157 (1997)

[5] Stew´enius, H., Schaffalitzky, F., Nist´er, D.: How hard is three-view triangulation really? In: Int. Conf. Computer Vision, Beijing, China, pp. 686–693 (2005)
[6] Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
[7] Triggs, W., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.: Bundle adjustmentfor structure from motion. In: Vision Algorithms: Theory and Practice, pp. 298–372. Springer, Heidelberg (2000)
[8] Boyd, S., Vanderberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
[9] Sturm, J.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optimization Methods and Software 11–12, 625–653 (1999)
[10] Kahl, F., Agarwal, S., Chandraker, M.K., Kriegman, D., Belongie, S.: Practical Global Optimization for Multiview Geometry. International Journal of Computer Vision 79(3), 271–284 (2008)

Erratum to: An Efficient Method for Target Extraction of Infrared Images

Ying Ling and Xingjin Mao

School of Computer Science Northwestern Polytechnical University
Shanxi, Xi'an, 710129, China
lybyp@nwpu.edu.cn, Maoxingj1985@gmail.com

F.L. Wang et al. (Eds.): AICI 2010, Part I, LNAI 6319, pp. 185–192, 2010.

DOI 10.1007/978-3-642-16530-6_59

The name of the first author was erroneously printed "Ying Ling". The correct spelling is "Ying Li".

The original online version for this chapter can be found at
http://dx.doi.org/10.1007/978-3-642-16530-6_23

Author Index

Ambraziunas, Martas I-230
Anbaroglu, Berk I-116

Bai, YunLong I-107
Bänfer, Oliver I-414
Benda, Zhou II-134
Bo, Hua I-397
Brezany, Peter I-97

Cai, Hegao I-292
Cai, Lei I-193
Cai, Nengbin I-321
Chang, Hongrui I-169
Chang, Tingjen II-374
Chen, Caikou I-331
Chen, Dongfeng II-350
Chen, Hua I-349
Chen, Peijun II-45
Chen, Taolue I-41
Chen, Wang I-145
Chen, Wei I-145
Chen, Xinyong I-273
Cheng, Ran II-293
Cheng, Tao I-116
Chu, Hua II-241

Dai, Gang I-97
Dai, Wensheng II-374
Deng, Nan I-321
Ding, Hao I-153
Ding, Hui I-432
Ding, Xiaoling I-145
Dong, Yunfeng II-114
Duan, Bobo I-311
Duan, Jing I-357

Ebert, Tobias I-414
El-Bakry, Hazem II-226

Fang, Gu II-326
Fu, Boxian II-168
Fu, Yili I-292
Fu, Zhenyong I-321
Fujimoto, Hideo II-64

Ganggang, Niu I-222
Gao, Chong I-161
Gao, Jun I-448
Gao, Junxiang II-358
Gao, Wei I-406
Gao, Weiwei I-389
Gao, Xiang II-106
Gao, Ying II-106
Geng, Xin II-275
Gong, Hongli I-440
Gu, Zhenghui II-71
Guo, Kehua I-373
Guo, Lei I-381

Hamada, Mohamed II-226
Han, Shizhong I-389
Hao, Fufei I-153
He, Dakuo I-169
He, Lin II-71
He, Tongneng II-45
He, Zhiwei I-10
Hedjazi, Seyyed Mahdi II-193
Horng, Ming-Huwi II-318
Hu, Di II-114
Hua, Cheng I-88
Huang, Dongmei I-301
Huang, Jing I-389
Huang, Pu I-331

Igarashi, Shigeru I-256
Ikeda, Akio II-79
Ikeda, Yasuwo I-256

Jia, Ji-Teng II-201
Jia, Lei II-58, II-275
Jia, Xiuping II-218
Jiang, Kaifei II-16
Jiang, Ting-Wei II-318
Jiang, Weijin I-145
Jiao, Jingshan II-350
Jiao, Wenpin I-132
Ji, Guoqing I-248
Ji, Jia II-185
Jie, Junjing I-132

Jingyao, Li I-222
Jung, Tzyy-Ping II-233

Kanehira, Ren II-64
Kao, Chao-Shun II-142
Kuang, Mingxia II-342
Kwok, Ngai Ming II-218, II-326

Lan, Yuqing II-342
Lei, Jingsheng I-49
Lei, Xiujuan II-98, II-267
Li, Feng II-218
Li, Honghui I-97
Li, Jie I-455
Li, Jing II-175
Li, Kai I-273
Li, Lijuan II-209
Li, Mengxin I-470
Li, Qian II-175
Li, Qiao I-239
Li, Qingshan II-241
Li, Shujin I-153
Li, Xiaorun I-422
Li, Xiaotian II-16
Li, Xiao-dong I-161
Li, Yanping I-432
Li, Ying I-440
Li, Yuanqing II-71
Li, Zhilao II-31
Li, Zhongxue II-249
Lian, Li-ting I-64
Liang, Hui II-87
Liang, Jinjin I-266
Liang, Yongchun I-1
Liang, Yun II-1
Lin, Xiao-Lin II-201
Lin, Zhengkui I-357
Ling, Ying I-185, E1
Liu, TongLin I-470
Liu, Jianxin II-9
Liu, Jinbiao I-10
Liu, Jun II-209
Liu, Li-Jen II-142
Liu, Sheng-dao I-64
Liu, Tao II-58
Liu, Wei I-311
Liu, Yanfang II-342
Liu, Yong II-358
Liu, Yuanyuan I-10
Liu, Zhifeng I-209
Liu, Zhijing I-479
Long, Chengjiang I-389
Lopata, Audrius I-230
Lu, Chen II-38
Lu, Chi-Jie II-374
Lu, Hongtao I-71, I-321
Lu, Jian I-41
Lu, Mingyu I-357
Luan, Jiahui II-38
Luo, Guiming I-25, I-201
Luo, Yuan I-455
Lv, Teng I-49, I-57

Ma, Qianzhi II-267
Mao, Shaojie II-241
Mao, Xingjin I-185, E1
Marjani, Samane Sadat II-193
Mei, Hong I-132
Min, Yang I-488
Minghua, Chen II-134
Mizutani, Tetsuya I-256

Nagamine, Takashi II-79
Nakamura, Masatoshi II-79
Nan, Deng I-71
Narita, Hirohisa II-64
Nelles, Oliver I-414
Nengbin, Cai I-71
Ng, Kwan-Chew II-175
Ning, Xiaozhou II-31

OuYang, Dantong I-18

Pan, Donghua I-124
Peng, Jinhua II-185
Peng, Yong II-124
Pu, Dongbing II-367

Qian, Yuntao I-365
Qin, Xiaoxia I-301
Qu, Kangkang II-31

Ren, Qian I-292
Ren, Zengqiang II-367

Sarosh, Ali II-114
Shen, Xiaojing I-301
Shen, Yinhe I-422
Shi, Chen-Xi II-175
Shi, Fengying II-9
Shi, MingYong II-87

Shi, Tingxun I-132
Shi, Yan-jun II-334
Shibasaki, Hiroshi II-79
Shio, Masayuki I-256
Shudong, Sun I-222
Song, Bao-wei I-33, II-310
Song, Qing I-455
Sugi, Takenao II-79
Sun, Bo I-209, I-349
Sun, Shanwu I-18
Sun, Yingjuan II-367
Sun, Zhen I-153
Suzuki, Tatsuo I-256

Tang, Diankui II-275
Tang, Jian II-38
Teng, Hong-fei II-334

Wang, Anlin II-16
Wang, Bei II-79
Wang, Dalong II-218, II-326
Wang, Hailan II-24
Wang, Lanzhou I-239
Wang, Lei II-241
Wang, Lifeng I-169
Wang, Lili I-177
Wang, Maoguang I-132
Wang, Nan I-18
Wang, Peng I-33, II-310
Wang, Shangping II-259
Wang, Shuguo I-292
Wang, Weihua I-479
Wang, Wenbin II-302
Wang, Xiaofeng II-259
Wang, Xike II-249
Wang, Xingyu II-79
Wang, Yijun II-233
Wang, Ying II-9
Wang, Yi-shou II-334
Wang, Yu-Jie II-142
Wang, Yu-Te II-233
Wang, Zhaoxia I-153
Wang, Zhiyong I-153
Wei, Shimin I-381
Wen, Chuanbo II-1
Wen, Guoguang I-339
Wöhrer, Alexander I-97
Wu, Chengdong I-470
Wu, Bin I-107
Wu, De I-266
Wu, Dongmei I-292
Wu, Jiansheng II-284
Wu, Xindong I-448

Xiao, Ailin I-177
Xiao, Chang-han I-64
Xiao, Jun II-9
Xie, Jiongyun I-71
Xie, Jun I-88
Xiong, Lu I-389
Xu, Guanghua I-88
Xu, Hua I-282
Xu, Jianhua I-282
Xu, Xiangzhong II-152
Xu, Xiaojun II-98
Xu, Yuhui I-145
Xue, Zhi-chun II-124

Yan, Ping I-49, I-57
Yan, Yunju II-31
Yang, Guangfei I-124
Yang, Jingyu I-331
Yang, Jun I-177
Yang, Ming-ming I-64
Yang, Nan I-273
Yang, Weiping II-64
Yao, Min II-293
Ye, Xiuchen I-273
Yi, Weiguo I-357
Yong, Longquan I-461
Yu, Fusheng II-160
Yu, Haibin I-10
Yu, Hongfei I-311
Yu, Hui II-209
Yu, Xiuli I-381
Yu, Yongguang I-339
Yu, Yongtao II-152
Yu, Zhuliang II-71
Yuan, Huai I-311
Yuan, Huang I-222
Yuan, Zhiyong I-389
Yue, Ben-xian II-334
Yue, Binnan II-16

Zhai, Yandong II-367
Zhang, Hao II-358
Zhang, Jian-ze I-161
Zhang, Lei II-350
Zhang, Linghua I-432
Zhang, Mingxin II-160
Zhang, Ning II-168

Zhang, Qing I-88
Zhang, Shi II-209
Zhang, Xibao II-52
Zhang, Xuexia I-193
Zhang, Yan I-97
Zhang, Ying I-470
Zhang, Yulai I-201
Zhao, Hong I-311
Zhao, Jianhui I-389
Zhao, Jianyu II-275
Zhao, Jun I-349
Zhao, Liaoying I-422
Zhao, Ran I-339
Zhao, Xinchao II-185, II-302
Zhao, Yandong II-24
Zhao, Yang I-25
Zhao, Yuanyuan I-169
Zhao, Zhizhong II-168
Zhao, Zhong-Qiu I-448
Zheng, Ping I-79
Zheng, Wenbin I-365
Zheng, Yi I-79
Zhong, Haiming II-249
Zhou, Conghua I-209
Zhou, Guo-hua I-64
Zhou, Tong II-367
Zhu, Hongbo I-248
Zhu, Qian I-124
Zhu, Qi-feng I-33, II-310
Zou, Cunwei I-455